Aus dem Institut für Trainingswissenschaft und Sportinformatik
der Deutschen Sporthochschule Köln
Geschäftsführender Leiter: Univ.-Prof. Dr. Daniel Memmert

Unterstützung der Leistungsdiagnostik im Leistungs- und Hochleistungssport durch maschinelles Lernen am Beispiel der Ausdauerdiagnostik

Von der Deutschen Sporthochschule Köln
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
(Dr. rer. nat.)

angenommene Dissertation

vorgelegt von

Benedikt Malecki
aus
Leverkusen

Köln 2021

Erster Gutachter:	Univ.-Prof. Dr. Dr. h.c. mult. Joachim Mester
Zweite Gutachterin:	Prof. Dr. Heide Faeskorn-Woyke
Vorsitzender des Promotionsausschusses:	Univ.-Prof. Dr. Mario Thevis
Datum der Disputation:	16.09.2021

Eidesstattliche Versicherung gem. § 7 Abs. 2 Nr. 4 und 5 der Promotionsordnung der Deutschen Sporthochschule Köln, 20.02.2013:

Hierdurch versichere ich:
Ich habe diese Arbeit selbständig und nur unter Benutzung der angegebenen Quellen und technischen Hilfen angefertigt; sie hat noch keiner anderen Stelle zur Prüfung vorgelegen. Wörtlich übernommene Textstellen, auch Einzelsätze oder Teile davon, sind als Zitate kenntlich gemacht worden.

Hierdurch erkläre ich, dass ich die „Leitlinien guter wissenschaftlicher Praxis“ der Deutschen Sporthochschule Köln eingehalten habe.

Datum, Unterschrift

Inhaltsverzeichnis

Abbildungsverzeichnis **iv**

Tabellenverzeichnis **vi**

Listingverzeichnis **vii**

Abkürzungsverzeichnis **viii**

1 Einleitung **1**

1.1 Problemstellung 1
1.2 Forschung und Entwicklung 5
1.3 Zielsetzung und Abgrenzung 11
1.4 Gliederung 13

2 Grundlagen **14**

2.1 Prozesse 14
2.1.1 Data-Warehouse-Prozess 14
2.1.2 Data Mining 19
2.2 Machine Learning 22
2.2.1 Unüberwachtes Lernen 22
2.2.2 Überwachtes Lernen 27
2.3 Ausdauer- und Laboratoriumsdiagnostik 33
2.3.1 Datenerhebung 33
2.3.2 Parameter 35

3 Material und Methodik **41**

3.1 Vorgehensweise 41
3.2 Datengrundlage 47
3.3 Software 49

4 Prototypisches System **51**

4.1 Anforderungen 51

4.2 Architektur 54
4.3 Umsetzung 57
4.3.1 Data-Server 57
4.3.2 CSV-Client 69
4.4 Anwendung 74

5 Exemplarische Ergebnisse: Beschreibung und Diskussion 80
5.1 Cluster und Gruppen 80
5.2 Einzelbetrachtung der Parameter 92
5.2.1 Zeit bis zum Abbruch 92
5.2.2 Im Verlauf: Anaerobe Schwelle V4 96
5.2.3 Relative maximale Sauerstoffaufnahme (peak) 99
5.2.4 Respiratorischer Quotient (peak) 104
5.2.5 Maximale Herzfrequenz 108
5.2.6 Blutlaktatkonzentration (peak) 112
5.2.7 Hämoglobin-Wert 116
5.3 Parameterübergreifende Regeln 121

6 Fazit und Ausblick 135
6.1 Fazit 135
6.2 Ausblick 139

Literatur 142

Anhang 154

A Konfigurationen des prototypischen Systems 154
A.1 Erstellen von Tabellen 154
A.1.1 Basisdatenbank 154
A.1.2 Ableitungsdatenbank 157
A.2 Datenintegration 160
A.2.1 Basisdatenbank 160
A.2.2 Ableitungsdatenbank 161

B Physiologische Parameter in Bezug auf die Zeit bis zum Abbruch **164**

B.1 Relative maximale Sauerstoffaufnahme (peak) 164

B.2 Respiratorischer Quotient (peak) 165

B.3 Maximale Herzfrequenz 167

B.4 Blutlaktatkonzentration (peak) 169

B.5 Hämoglobin-Wert 171

C Stabilitätsanalyse des Clusterings **173**

Zusammenfassung **178**

Abstract **180**

Abbildungsverzeichnis

1 Data-Warehouse-Prozess 15
2 Modifizierte Referenzarchitektur 17
3 CRoss Industry Standard Process for Data Mining 20
4 Konzeptteil 1 der Vorgehensweise 42
5 Konzeptteil 2 der Vorgehensweise 42
6 Konzeptteil 3 der Vorgehensweise 43
7 Komponentendiagramm *Gesamtüberblick* 54
8 Aktivitätsdiagramm *Tabelle erstellen* 59
9 Klassendiagramm *Tabelle erstellen* 60
10 Aktivitätsdiagramm *Integration eines einzelnen Datensatzes* . 63
11 Aktivitätsdiagramm *Integration von multiplen Datensätzen* . . 64
12 Aktivitätsdiagramm *Zeilenweise Integration von multiplen Datensätzen* 66
13 Klassendiagramm *Daten-Integration* 68
14 Aktivitätsdiagramm *CSV-Upload* 71
15 Klassendiagramm *CSV-Client* 73
16 Systemanwendung für das Erstellen der Tabellen innerhalb der Basisdatenbank 75
17 Systemanwendung für die Datenintegration und -aggregation . 76
18 Systemanwendung für das Erstellen der Tabellen innerhalb der Ableitungsdatenbank 78
19 Systemanwendung für den Upload 79
20 Graphische Darstellung des Dendogramms 81
21 Mittelwerte der Cluster für die untersuchten Parameter 83
22 Balkendiagramm *Geschlechterverteilung* 84
23 Balkendiagramm *Altersverteilung von Cluster low* 85
24 Balkendiagramm *Altersverteilung von Cluster high* 86
25 Balkendiagramm *Sportartenverteilung* 87
26 Boxplots *Alter nach Gruppen* 89
27 Boxplots *tlim nach Gruppen* 93
28 Boxplots *V4 nach Gruppen* 96

29 Boxplots $rVO_{2_{peak}}$ *nach Gruppen* 100
30 Boxplots RQ_{peak} *nach Gruppen* 105
31 Boxplots Hf_{max} *nach Gruppen* 109
32 Boxplots Lak_{peak} *nach Gruppen* 113
33 Boxplots *Hb nach Gruppen* 117
34 Decision Tree der physiologischen Parameter 121
35 Ausprägungskombinationen der $\mathrm{Lak}_{\mathrm{peak}}$ und der tlim nach Regeln 127
36 Extreme Ausprägungskombinationen der $\mathrm{Lak}_{\mathrm{peak}}$ und der tlim nach Regeln 129
37 Extreme Parameterausprägungen einzelner Individuen 131
38 Dashboard 140
39 Wertekombinationen der $\mathrm{rVO}_{2_{\mathrm{peak}}}$ und der tlim 164
40 Wertekombinationen des $\mathrm{RQ}_{\mathrm{peak}}$ und der tlim 166
41 Wertekombinationen der $\mathrm{Hf}_{\mathrm{max}}$ und der tlim 168
42 Wertekombinationen der $\mathrm{Lak}_{\mathrm{peak}}$ und der tlim 170
43 Wertekombinationen des Hbs und der tlim 171
44 Graphische Darstellung des Dendogramms aus der Stabilitätsanalyse 173
45 Parametermittelwerte der Cluster aus der Stabilitätsanalyse . 175

Tabellenverzeichnis

1 Parameter der Ausdauer- und Laboratoriumsdiagnostik 35
2 Datengrundlage 47
3 Parametermittelwerte von Cluster low 81
4 Parametermittelwerte von Cluster high 82
5 Statistische Kenngrößen des Alters [Jahre] 89
6 Statistische Kenngrößen der tlim [min] 92
7 Statistische Kenngrößen der V4 [m/s] 97
8 Statistische Kenngrößen der $rVO_{2_{peak}}$ [ml/kg/min] 99
9 Statistische Kenngrößen des RQ_{peak} 104
10 Statistische Kenngrößen der Hf_{max} [S/min] 108
11 Statistische Kenngrößen der Lak_{peak} [mmol/l] 112
12 Statistische Kenngrößen des Hb [g/dl] 118
13 Regeln für die physiologischen Parameter 124
14 Individuenwerte für Alter, tlim und V4 131
15 Individuenwerte für $rVO_{2_{peak}}$, Hb und Hf_{max} 132
16 Individuenwerte für Lak_{peak} und RQ_{peak} 132
17 Parametermittelwerte von Cluster low der Stabilitätsanalyse . 174
18 Parametermittelwerte von Cluster high der Stabilitätsanalyse . 175

Listings

1 JSON-Format für das Anlegen einer Tabelle 57
2 JSON-Format für die Integration eines einzelnen Datensatzes . 62
3 JSON-Format für die Integration multipler Datensätze 62
4 JSON-Format von Erfolgs- und Fehlermeldungen 65
5 Konfiguration des CSV-Uploads 70
6 Port- und Datenbankkonfiguration der Data-Server-Instanz Integration 74
7 Port- und Datenbankkonfiguration der Data-Server-Instanz Delivery 75
8 Konfiguration der Tabelle master_data 154
9 Konfiguration der Tabelle endurance 155
10 Konfiguration der Tabelle lab 156
11 Konfiguration der Tabelle endurance_lab 157
12 Konfiguration der Tabelle endurance_lab_with_cluster_ids . 158
13 Konfiguration der Stammdatenintegration 160
14 Konfiguration der Integration von Ausdauerdiagnostikdaten . 160
15 Konfiguration der Integration von Laboratoriumsdiagnostikdaten 161
16 Integrationskonfiguration der aggregierten Datensätze 161
17 Integrationskonfiguration der modifizierten Datensätze 162

Abkürzungsverzeichnis

API Application Programming Interface

Aus Ausreißer

BI Business Intelligence

CDSS Clinical Decision Support System

CRISP-DM CRoss Industry Standard Process for Data Mining

CSV comma-separated values

CWR constant work rate

DAO Data Access Object

DFB Deutscher Fußballbund

momentum Deutsches Zentrum für Leistungssport Köln

DSHS Deutsche Sporthochschule Köln

DSS Decision Support System

eAkte elektronische Athletenakte

EDTA Ethylendiamintetraessigsäure

ETL Extraktion-Transformation-Laden

ExHR exercising heart rate

GPS Global Positioning System

GRNN General Regression Neural Network

Hb Hämoglobin-Wert

Hf_{max} maximale Herzfrequenz

HMV Herzminutenvolumen

HTTP Hypertext Transfer Protocol

IBM International Business Machines Corporation

IDE Integrated Development Environment

ID Identifier

IQR Inter Quartil Range

JSON JavaScript Object Notation

JAAI Just Add AI

KI Künstliche Intelligenz

Lak_{peak} Blutlaktatkonzentration (peak)

Ltd. private limited company

LDU leistungsdiagnostische Untersuchung

Max Maximum

MCT Monocarbocylat-Transporter

Med Median

Min Minimum

ML Machine Learning

MLP Multilayer Perceptron

MLR Multiple Linear Regression

MLSS maximales Laktat-steady-state

MW Arithmetischer Mittelwert

O_2 Sauerstoff

PaaS Platform as a Service

Q-PFA Perceived Functional Ability

POJO Plain Old Java Object

$\mathbf{RQ_{peak}}$ Respiratorischer Quotient (peak)

SD Standardabweichung

SD-ID Stammdaten-Identifier

SMU sportmedizinische Untersuchung

SQL Structured Query Language

SVM Support Vector Machine

Q1 Quartil 1

Q3 Quartil 3

tlim Zeit bis zum Abbruch

URL Uniform Resource Locator

V4 Im Verlauf: Anaerobe Schwelle V4

$\mathbf{rVO_{2peak}}$ relative maximale Sauerstoffaufnahme (peak)

$\mathbf{rVO_{2max}}$ relative maximale Sauerstoffaufnahme

XML Extensible Markup Language

1 Einleitung

Innerhalb des vorliegenden Kapitels wird das in dieser Arbeit behandelte Problem herausgearbeitet, um anschließend einen Überblick über den Stand der aktuellen Forschung und Entwicklungen zu geben. Diesem Überblick folgt die Zielsetzung und Abgrenzung der Arbeit. Den Abschluss bildet die Gliederung, die dem Leser einen Überblick über die Inhalte dieser Arbeit vermittelt.

1.1 Problemstellung

Im Sportgeschehen streben Aktive danach, die Grenzen der menschlichen Leistungsfähigkeit bei der Jagd nach neuen Rekorden sowohl in Team- als auch in Einzeldisziplinen immer weiter zu verschieben.

Für die Realisierung solcher Extremleistungen arbeiten unter anderem auch Trainer sowie Sportwissenschaftler an der Optimierung dieser Aktiven (Bellinger & Brennan, 2018). So wird in der Sportwissenschaft beispielsweise im Leistungsfußball bereits seit längerer Zeit an der Typisierung von Spielern und an Controllingverfahren geforscht, um entsprechende Informationen für Training und Wettkampf zur Verfügung stellen zu können (Broich, 2009).
Dabei werden Sportwissenschaftler und Trainer durch datenhaltende Systeme wie beispielsweise die elektronische Athletenakte (eAkte) an der Deutschen Sporthochschule Köln (DSHS) unterstützt. Dieses System speichert bis zu 2150 Attribute pro Athletin oder Athlet (Nöll, 2009).
Weitere Unterstützung bei der Aggregation einer solchen Anzahl an Attributen sowie der Analyse einer solchen Datenflut insgesamt versprechen hier Verfahren, die unter dem Schlagwort Künstliche Intelligenz (KI) immer wieder im aktuellen Sportgeschehen auftauchen. Bereits bei der Suche nach Talenten und damit zukünftigen Leistungsträgern im Sport kommen Systeme basierend auf KI zum Einsatz, wie dies beispielsweise bei dem Fußballbundesligisten SV Werder Bremen der Fall ist (Schnor, 2018). Bei diesem unterstützt eine Software des Startups Just Add AI (JAAI) Fußballscouts dadurch, dass sie deren Berichte, die in unstrukturierter Form in einer Datenbank gespei-

chert sind, ausliest, alle relevanten Information extrahiert und diese Informationen als Profile zu einzelnen Spielern zur Verfügung stellt (Schnor, 2018). Die Grundlage dieser Software bildet wiederum die Software Watson[1] des Unternehmens IBM. Für die Erstellung der Profile werden unter anderem auch Informationen aus Datenquellen wie sozialen Netzwerken verwendet (Schnor, 2018; Göbel, 2019; Moeser, 2019). Solche Quellen sind jedoch teilweise manipulierbar (Eckert, Hurtz, Müller-Hansen & Wormer, 2019). Somit kann hier die Frage nach der Aussagekraft von Spielerprofilen gestellt werden, deren Datenbasis mitunter auf Einträgen aus sozialen Medien beruht.

Jedoch nicht nur Firmen wie JAAI proklamieren den Einsatz von KI im Sportgeschehen. Auch im Deutschen Fußballbund (DFB) erhofft man sich Fortschritte im Fußball durch KI (Ahrens, 2018). Aber nicht nur im deutschen Fußball sollen Algorithmen der KI zur Trainingsunterstützung eingesetzt werden, wie die Zusammenarbeit des FC Liverpool mit der IT-Firma Arcronis zeigt (Köpf, 2019).

Auch für die Planung des taktischen Einsatzes von Spielern existieren Überlegungen, entsprechende Daten wie beispielsweise Essverhalten, Herzfrequenz sowie ihr Schlafverhalten zu erheben (Boeselager, 2018). Darüber hinaus existieren bereits Versuche, die Leistung einzelner Spieler in Wettkampfbegegnungen mess- und berechenbar zu machen (Schlichtmeier, 2019).

Solche Taktikplanungen beschränken sich jedoch nicht nur auf einzelne Spielereinsätze, sondern umfassen auch ein gesamtes Team. So werden neben individuellen physiologischen Ausprägungen auch technische Fähigkeiten und Teamaufstellungen sowie Wechselwirkungen zwischen diesen Faktoren betrachtet (Rein & Memmert, 2016). Dabei kommen unterschiedlichste Verfahren wie beispielsweise geometrische, geostatistische oder stochastische Modelle so wie auch Regressionsmodelle zum Einsatz (Link, 2018).

Wie schwer jedoch die zuverlässige Anwendung von Verfahren der KI im aktuellen Sportgeschen ist, zeigen Beispiele wie die misslungene Vorhersage des

[1]Watson ist gemäß der Autoren Rauch und Litzel (2016) die Antwort der International Business Machines Corporation (IBM) auf Cognitive Computing. Die Software findet laut Gliozzo et al. (2017) Anwendung beim Erforschen von Daten, beim Finden neuer Korrelationen und neuer Zusammenhänge, um neue Lösungen bereitstellen zu können. Die APIs von Watson können als Platform as a Service (PaaS) eingebunden werden.

Fußballweltmeisters für 2018 (Groll, Ley, Schauberger & Eetvelde, 2018).
Im aktuellen Zeitgeschehen existieren deshalb auch Stimmen, die die Anwendung von KI zum jetzigen Zeitpunkt kritisch hinterfragen. So sind nach Dworschak (2018) Systeme, die unter dem Begriff der KI verwendet werden, nur in der Lage, eng umrissene Aufgaben zu erfüllen. Der Arbeitsaufwand, bis entsprechende Systeme für umfassende Aufgaben akzeptable Ergebnisse liefern, ist daher immens (Dworschak, 2018).

Eine eng umrissene Aufgabe für ein System, das auf KI basiert, könnte jedoch teilweise in der Unterstützung des Betreuungspersonals bei der Trainingsplanung liegen. Als Grundlage für ein solches System können die Daten von standardisierten Diagnostiken verwendet werden, die in verschiedenen Sportarten erhoben werden, um eine Bestandserhebung des aktuellen Leistungsstandes eines Individuums zu erhalten. Eine solche Diagnostik ist beispielsweise die Spiroergometrie für die Bestimmung der Ausdauer, welche essentiell für Höchstleistungen in Sportarten wie beispielsweise Boxen, Fußball oder Handball ist (Hohmann, Lames, Letzelter & Pfeiffer, 2020).

Aufgrund der Datenauswertung aus den durchgeführten Untersuchungen der Spiroergometrie einzelner Individuen eines Kaders sollte das betreuende Personal eines solchen Kaders in die Lage versetzt werden, ein möglichst individuelles[2] und effektives Training im Bereich der Ausdauer in Hinblick auf eine Leistungserhaltung oder sogar eine mögliche Leistungssteigerung anbieten und durchführen zu können.
Für die strukturelle Übersicht sowohl auf leistungsdiagnostischer als auch physiologischer Ebene stoßen statistische Verfahren wie die Berechnung des arithmetischen Mittelwertes oder die Standardabweichung an ihre Grenzen, um dem Betreuungspersonal einen Überblick über einen Leistungskader und dessen Individuen zu vermitteln.
Hier stellt sich die Frage, auf welche Art und Weise individualisierte Informationen zur Verfügung gestellt werden können. Eine Antwort auf diese Frage geben im klinischen Umfeld innerhalb der personalisierten Medizin Clini-

[2]in Bezug auf ein einzelnes Individuum innerhalb des Kaders

cal Decision Support Systems (CDSSs) (Sutton et al., 2020) sowie digitale Zwillinge (Björnsson et al., 2019; *About the Initiative*, o. J.). Ein solch umfassender und personalisierter Ansatz, wie er teilweise in der personalisierten Medizin existiert, ist in der Sportwissenschaft für die Bereiche des Trainings und der sportmedizinischen Betreuung aktuell jedoch nicht vorhanden.

1.2 Forschung und Entwicklung

Das vorliegende Unterkapitel betrachtet die Notwendigkeit und Existenz von Decision Support Systems (DSSs) innerhalb der Sportwissenschaft. In diesem Kontext werden zunächst freie Business Intelligence (BI)-Systeme und Datenintegrationswerkzeuge aufgeführt, um anschließend einen kurzen Umriss zu dem Begriff der KI zu geben. Im Anschluss werden kommerzielle Anbieter mit ihren Softwarelösungen betrachtet, welche Entscheidungsunterstützungen im Sportbereich – vor allem im Profifußball – bewerben. Zuletzt werden DSSs sowie andere Veröffentlichungen in der Sportwissenschaft betrachtet, welche mit Hilfe von Machine Learning (ML) Entscheidungshilfen bei der Planung von Ausdauertraining geben können.

Auf dem freien Markt existieren diverse kommerzielle Anbieter, die DSSs im Bereich des Sports bewerben. Bei näherer Betrachtung dieser Anbieter liegt deren Fokus größtenteils auf dem Geschehen des nationalen und internationalen Profifußballs.
So bietet beispielsweise der Betreiber Stats Perform (*Revolutionise Sport Through AI*, o. J.) sowohl Analysen von Sportdaten als auch Entscheidungsunterstützung mit Hilfe von ML an. Ein anderer Anbieter namens Soccerment (*Soccerment*, o. J.) verspricht durch den Dienst Soccerment Analytics (*Soccerment Analytics*, o. J.) ebenfalls Leistungsanalysen zu Fußballspielern und Fußballteams. Das Unternehmen StatsBomb bewirbt mit seiner Plattform Spielanalysen sowie Spielercharakteristiken (*StatsBomb is built by football experts*, o. J.). Auch die private limited company (Ltd.) Soccerlogic (*Soccerlogic*, o. J.) stellt Hilfen bei der Entscheidungsfindung sowohl bei der Spielvorbereitung als auch beim Scouting von Spielern in Aussicht.
In der Sportwissenschaft werden auch im aktuellen Zeitgeschehen informatische Ansätze für taktische Analysen von Fußballspielen im professionellen Bereich untersucht. Zu solchen Ansätzen zählen beispielsweise die Vergleiche von Parametern wie inter-player coordination oder inter-team coordination vor kritischen Geschehen wie Torerfolgen (Memmert, Lemmink & Sampaio, 2016).

Die Grundlage dieser Systeme für eine Entscheidung basiert teilweise auf Tracking-Sytemen, die Daten aus dem Global Positioning System (GPS) oder Video-Systemen auswerten. Diese Grundlage erschwert jedoch eine Vergleichbarkeit der Entscheidungen verschiedener Systeme, da diese DSSs teilweise unterschiedliche Tracking-Systeme verwenden (Linke, Link & Lames, 2018).

Auch innerhalb der Sportwissenschaft ist gemäß Ward, Windt und Kempton (2019) die Notwendigkeit für Prozesse zur Entscheidungsunterstützung erkannt worden, um wissenschaftliche Erkenntnisse und Einsichten in der Sportpraxis anwendbar machen zu können. Sportwissenschaftlerinnen und Sportwissenschaftler sollten demnach den Entscheidungsprozess innerhalb ihrer Organisation auf der Basis eines Decision Support Models mitgestalten. Ein solches Modell umfasst gemäß der Autoren die folgenden Bereiche:

1. Data collection and organisation
2. Analytic models to drive insight
3. Interface and communication of information

Entsprechende BI-Systeme für die Organisation von Daten (data organisation nach Ward et al. (2019)) sind als freie Lösungen verfügbar. King (2019b) nennt die folgenden:

- BIRT: Offene Standards, Integration verschiedenster Datenquellen, Generierung von Datenvisualisierungen und Reports (*What is BIRT?*, o. J.)
- Helical Insight CE: Einbinden einzelner oder multipler Datenquellen (*Helical Insight*, o. J.)
- KNIME Analytics Platform: Analytics Plattform mit Einbindung verschiedenster Datenquellen wie Textdateien und Datenbanken (*KNIME Analytics Platform*, o. J.)
- Metabase: Dashboards mit Benachrichtigungen bei Änderung in Daten (*Metabase*, 2020)

Beim Sammeln der Daten (data collection) ist nach Ward et al. (2019) die Integration der Daten entscheidend, um unter anderem Daten Silos innerhalb der Organisation zu vermeiden. King (2019a) gibt einen Überblick über die folgenden freien Systeme zur Datenintegration mit den jeweiligen Funktionalitäten:

- Apache Kafka: Streamen von Records (*Apache Kafka® is a distributed streaming platform. What exactly does that mean?*, o. J.)
- Apache NiFi: Steuerung von Datenflüssen zwischen Systemen (*Apache NiFi Overview*, 2020)
- CloverDX: Integration und Transformation von Daten (*CloverDX*, 2020)
- Jaspersoft® ETL: Extraktion von Daten aus transaktionalen Systemen (*Jaspersoft® ETL*, o. J.)
- KETL™: Extraktion-Transformation-Laden (ETL) und Scheduling für Datenintegration (*KETL*, 2015)
- Kettle: ETL, basierend auf Metadaten-Ansatz (*Data Integration - Kettle*, 2017)
- Talend Open Studio for Data Integration: Free open source Apache licence, ETL, Versionierung, Connectoren zu RDBMS (*Open Studio for Data Integration*, o. J.)
- Scriptella: Open source Tool für ETL und Skripting (*Welcome to Scriptella ETL Project*, o. J.)
- Apatar: Open source Datenintegrationssoftware, Anwendung ohne Programmier-Kenntnisse (*What Is Apatar Open Source Data Integration?*, o. J.)

Der Einsatz von DSSs zur Unterstützung von Sport-Teams findet gemäß Robertson, Bartlett und Gastin (2017) aktuell bereits in Form von Ampelsystemen im Hochleistungssport statt. Solche Systeme entsprechen in ihrer Ausprägung dem dritten Teil des Decision Support Models nach Ward et al.

(2019) (interface and communication of information). Die Autoren weisen jedoch darauf hin, dass zukünftig auch Ansätze aus dem ML und damit der KI in Bezug auf Entscheidungsfindung evaluiert werden müssen.
Erste Studien wie beispielsweise die von Rommers et al. (2020) setzen Algorithmen aus dem ML ein, um Verletzungen von Elite-Jugendspielern im Fußball vorhersagen und somit vermeiden zu können.
Solche Ansätze, wie Robertson et al. (2017) sie fordern und Rommers et al. (2020) sie bereits anwenden, entsprechen den analytischen Modellen (analytic models to drive insight), die von Ward et al. (2019) proklamiert werden.

Das Forschungsfeld der KI umfasst die Disziplinen der Philosophie, Mathematik, Wirtschafts- und Neurowissenschaft, Psychologie, Technischen Informatik, Regelungstheorie und Kybernetik sowie Linguistik (Russel & Norvig, 2012). Deru und Ndiaye (2019) betrachten mit dem ML nur ein Teilgebiet der von Russel und Norvig (2012) beschriebenen Felder. Darüber hinaus bezeichnen Deru und Ndiaye (2019) Deep-Learning als eine spezielle Form von ML; Kaplan und Haenlein sehen in ML explizit die Basis für Deep-Learning-Verfahren (Kaplan & Haenlein, 2019). Im weiteren Verlauf dieser Arbeit werden ML-Verfahren angewendet.
Bei den ML-Verfahren ist ein tiefergehendes Verständnis der zu untersuchenden Daten notwendig, wie Deru und Ndiaye (2019) schreiben. Für die Erstellung eines ML-Modells muss zunächst eine Vorverarbeitung der Daten stattfinden. Im Anschluss daran hat eine Merkmalsextraktion und gegebenenfalls eine -reduktion zu erfolgen. Erst dann kann die gewünschte Modellentwicklung stattfinden.
Deep-Learning-Verfahren bilden hingegen End-to-End-Lernsysteme ab, bei denen auf eine explizite Merkmalsextraktion und Problemdekomposition verzichtet wird. Es findet somit eine direkte Abbildung der Eingabemuster auf die angestrebte Ausgabe statt. Anwendung finden Deep-Learning-Verfahren vor allem in der Sprachverarbeitung sowie bei automatischen Sprachübersetzungen, bei Visual Computing, autonomem Fahren, der Interpretation radiologischer Bilder in der Medizin, in der Finanzwelt und in der Energiewirtschaft, so die Autoren.

Verfahren der KI werden durch Cloud-Dienste wie beispielsweise Google Cloud (*AI and machine learning products*, o. J.) oder Amazon Web Services (*Machine Learning in AWS*, o. J.) angeboten, können jedoch auch offline durch die Verwendung von Frameworks beziehungsweise Bibliotheken wie Tensor-Flow (Abadi et al., 2015), Keras (Chollet et al., 2015) oder Scikit-Learn (Pedregosa et al., 2011) genutzt werden.

Um einen Überblick über aktuelle Veröffentlichungen in der Sportwissenschaft zu erhalten, die analytische Modelle zur Entscheidungsunterstützung basierend auf ML verwenden, wurde eine Literaturrecherche in den Recherchesystemen PubMed, ViFa:Sport, SURF, SPONET und SPORTDiscus mit den Suchbegriffen *machine learning* und *endurance* vorgenommen, die durch eine Und-Verknüpfung miteinander kombiniert wurden.
Abut und Akay (2015) geben einen Überblick über Veröffentlichungen mit Modellen zur Vorhersage der maximalen Sauerstoffaufnahme ($VO_{2_{max}}$). Laut der Autoren dienen diese Modelle dazu, in Bezug auf Zeit und Material aufwendige Testverfahren zu ersetzen. Die Modelle wurden in Studien mit Individuen unterschiedlichen Alters, Geschlechts und Fitnessstatus erstellt. Für die Erstellungen kamen laut Abut und Akay (2015) die Verfahren Support Vector Machine (SVM), Multilayer Perceptron (MLP), General Regression Neural Network (GRNN) und Multiple Linear Regression (MLR) zum Einsatz.
In ihrer Veröffentlichung beschreiben Abut, Akay und George (2016) die Verwendung von SVMs, um eindeutige Prädiktoren für die relative maximale Sauerstoffaufnahme ($rVO_{2_{max}}$) zu finden. Als vorhersagende Variablen wurden das Geschlecht, das Alter, die maximale Herzfrequenz, die submaximale Endgeschwindigkeit des Stufentests sowie die Perceived Functional Ability (Q-PFA) als am aussagekräftigsten befunden.
Neben Vorhersagemodellen für die maximale Sauerstoffaufnahme wurde bei der Literaturrecherche auch ein Paper identifiziert, das basierend auf einer exponentiellen Regression die Blutlaktatkonzentration bei niedriger, moderater und hoher constant work rate (CWR) vorhersagt (Huang et al., 2019). Neben der maximalen Herzfrequenz (Hf_{max}) beeinflussen nach Aussage der Auto-

ren auch respiratorische Variablen die Blutlaktatkonzentration bei niedriger und moderater CWR. Bei hoher CWR hat die exercising heart rate (ExHR) einen signifikanten Einfluss.
Eine weitere Veröffentlichung untersucht den Zusammenhang zwischen Anthropometrie und der Ausdauerleistungsfähigkeit von Wettkampfradfahrern (van der Zwaard, de Ruiter, Jaspers & de Koning, 2019). Die Autoren berichten, dass auf der Basis des k-means Clustering-Algorithmus drei anthropometrische Cluster gefunden werden konnten, die teilweise einen solchen Zusammenhang bestätigen.

Die vorliegenden Betrachtungen dieses Unterkapitels zeigen, dass DSSs im sportwissenschaftlichen Bereich gefordert werden und bereits Modellansätze auf der Basis von ML zur Entscheidungsfindung in einzelnen sportwissenschaftlichen Disziplinen existieren. Im Bereich des Ausdauertrainings existieren gemäß der vorliegenden Recherche jedoch noch keine Ansätze, die einen ganzheitlichen Prozess von der Datenerhebung bis zur -analyse umfassen und die bei der Planung des Ausdauertrainings von Individuen eines Kaders zur Unterstützung gezielt herangezogen werden können.

1.3 Zielsetzung und Abgrenzung

Das Ziel dieser Arbeit besteht in der konkreten Realisierung der beiden Bereiche *data collection and organisation* und *analytic models to drive insight* aus dem Decision Support Model[3] nach Ward et al. (2019), um sie bei der Entscheidungsfindung beim Ausdauertraining unterstützend heranzuziehen. Für die Erreichung dieses Ziels werden im Folgenden zwei Teilziele definiert, die den beiden oben aufgeführten Bereichen des Decision Support Models entsprechen.

Für den Bereich von data collection and organisation und damit dem ersten Teilziel ist im Rahmen dieser Arbeit ein System zu verwenden, das der Integration und Persistierung der zu analysierenden Daten dient.
Die Verwendung von großen und komplexen Systemen dieser Art ist im wissenschaftlichen Umfeld in Hinblick auf den Einarbeitungsaufwand sowie die Anschaffungs- und Betriebskosten für kleinere Projekte, wie beispielsweise Promotionen, nicht dauerhaft leistbar. Aus diesem Grund erscheint die Auswahl eines kleinen, leichtgewichtigen Systems, das ohne Programmierkenntnisse in seiner Datenstruktur erweitert werden kann, sinnvoll zu sein. Darüber hinaus erscheint auch eine potentielle Erweiterbarkeit in Hinblick auf seine Funktionalität sinnvoll.
Aus diesen Gründen wird innerhalb dieser Arbeit eine Applikation als Server-/Client-Architektur implementiert, die die Anforderungen von Datenintegration und -persistierung im Rahmen dieser Arbeit erfüllt.

Das zweite Teilziel (analytic models to drive insight) ist mit Hilfe von ML zu realisieren. Dabei ist ein Modell zu entwickeln, welches zum einen eine Einordnung von Individuen nach Leistungsparametern im Ausdauerbereich ermöglicht, zum anderen Aufschlüsse über die konkreten Ausprägungen der physiologischen Strukturen liefert, die für die Erbringung einer konkreten Leistung im Ausdauerbereich notwendig sind.

[3] Siehe auch Kapitel 1.2 ab S. 5.

Das Ziel dieser Arbeit ist somit nicht die Entwicklung eines Systems, das das gesamte Vorgehen von der Datenintegration bis hin zur -auswertung ohne Medienbrüche abbildet. Auch der Bereich *interface and communication of information* des Decision Support Models wird innerhalb dieser Arbeit nicht in einem System abgebildet. Vielmehr dient die vorliegende Arbeit selbst dem Zweck, die gewonnenen Informationen zu kommunizieren.

1.4 Gliederung

In Kapitel zwei ab S. 14 werden die Grundlagen dieser Arbeit beschrieben. Zunächst werden der Data-Warehouse-Prozess und der CRoss Industry Standard Process for Data Mining (CRISP-DM) vorgestellt. Anschließend werden die beiden Bereiche unüberwachtes und überwachtes Lernen des ML ausgeführt. Den Abschluss bilden eine Beschreibung des Ursprungs der in dieser Arbeit verwendeten Daten sowie eine kurze Übersicht über die in dieser Arbeit betrachteten spiroergometrischen und hämatologischen Parameter.

Das dritte Kapitel ab S. 41 umfasst das methodische Vorgehen, die verwendete Software sowie verwendete Software-Bibliotheken und einen Überblick über die in dieser Arbeit analysierten Datensätze.

Innerhalb des vierten Kapitels ab S. 51 werden die Systemanforderungen aufgeführt und die Architektur sowie die Umsetzung des Systems beschrieben.

Kapitel fünf ab S. 80 enthält die exemplarischen Ergebnisse dieser Arbeit. Zunächst werden die Entstehung der Leistungscluster und die aus diesen abgeleiteten Gruppen beschrieben. Anschließend werden die mit Hilfe des Clusterings erstellten Gruppen für jeden Parameter mit Hilfe von Boxplots deskriptiv betrachtet und diskutiert. Zum Abschluss dieses Kapitels werden unter Anwendung eines Decision Trees Regeln aufgestellt und diese im sportwissenschaftlichen Kontext diskutiert.

Den Abschluss dieser Arbeit bildet das sechste Kapitel ab S. 135 mit einer Zusammenfassung der Erkenntnisse aus der vorliegenden Arbeit sowie einem Ausblick für die Anwendung des implementierten Systems sowie des erstellten analytischen Modells.

2 Grundlagen

Das vorliegenden Kapitel beschreibt die Grundlagen dieser Arbeit. Zunächst werden die grundlegenden Prozesse aufgeführt. Im Anschluss daran wird ML betrachtet. Abschließend wird ein Überblick über die analysierten Parameter gegeben.

2.1 Prozesse

In diesem Unterkapitel werden sowohl der Data-Warehouse-Prozess als auch der CRISP-DM beschrieben.

2.1.1 Data-Warehouse-Prozess

Der Data-Warehouse-Prozess, auch als Data Warehousing bezeichnet (Bauer & Günzel, 2013), richtet sich an Entscheidungsträger und Data-Mining-Spezialisten (Lusti, 2002) und kann folgendermaßen definiert werden.

> *„Der Data-Warehouse-Prozess umfasst alle Schritte des Datenbeschaffungsprozesses, das Speichern und Auswerten der Daten.“* (Bauer & Günzel, 2013, S. 616)

Somit kann der Data-Warehouse-Prozess wiederum in den Prozess der *Datenbeschaffung* mit den aufeinander aufbauenden Phasen *Extraktion*, *Transformation* und *Laden* sowie den Prozess der *Auswertung* unterteilt werden, wie in Abb. 1 auf S. 15 zu sehen ist. Dabei setzt die Analyse die Datenbeschaffung voraus. Die Analyse kann beispielsweise durch *Data Mining* erfolgen.

Für die Umsetzung des Data-Warehouse-Prozesses wird ein Data-Warehouse-System benötigt (Bauer & Günzel, 2013). Einer der Ursprünge eines solchen Systems ist auf Inmon mit der folgenden Definition zurückzuführen.

> *„A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions.“* (Inmon, 2002, S. 31)

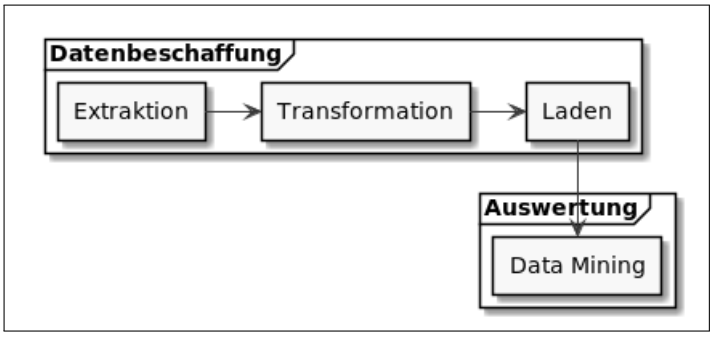

Abbildung 1: Data-Warehouse-Prozess mit Subprozessen (Quelle: Eigene Darstellung in Anlehnung an die Definition von Bauer und Günzel (2013, S. 615-616))

Nach dieser Definition sind Daten innerhalb eines Data Warehouse *fachorientiert* (engl. subject-oriented), *integriert* (engl. integrated), *beständig* (engl. nonvolatile) und *zeitvariant* (time-variant).
Laut Inmon (2002) spiegeln die Daten innerhalb eines Data Warehouse somit die Fachorientierung der jeweiligen Bereiche einer Organisation wider. Als wichtigstes Charakteristikum ist dabei aus seiner Sicht die integrierte Datenbasis eines Data Warehouse zu bezeichnen. In diese Basis fließen Daten aus unterschliedlichen Quellen der operativen Umgebungen ein. So bezeichnet er die Daten als beständig in dem Sinne, dass ein Zugriff regulär nur lesend erfolgt. Daten, die einmal in das Data Warehouse eingetragen sind, werden nach Meinung von Inmon regulär nicht manipuliert. Um die Daten des Weiteren über einen langen Zeitraum verfolgen zu können, wird zum Zwecke der Zeitvarianz jede Datenzeile mit einer Zeitmarkierung versehen. Eine solche Markierung kann nach Inmon sowohl in Form eines Zeitstempels als auch nur eines Datums erfolgen.
Bauer und Günzel geben darüber hinaus eine konkretere und eher technisch ausgerichtete Definition eines Data-Warehouse-Systems, in der jedoch auch eine integrierte Datenbasis zentraler Bestandteil ist.

> *„Ein Data-Warehouse-System ist ein physisches Informationssystem, das eine integrierte Sicht auf beliebige Daten zu Auswertungszwecken ermöglicht.“* (Bauer & Günzel, 2013, S. 8)

Diese Definition beinhaltet ebenso wie die von Inmon eine integrierte Sicht auf Daten. Ein Data Warehouse lässt sich somit anhand eines Datenbanksystems realisieren.

Der modifizierten Referenzarchitektur in Abb. 2 auf S. 17 kann der schematische Aufbau eines Data-Warehouse-Systems entnommen werden. Dieser umfasst *Externe Systeme* und das eigentliche *Data-Warehouse-System*. Externe Systeme enthalten *Datenquellen* oder spiegeln diese wider. Das Data-Warehouse-System selbst ist wiederum in einen *Integrations-* und einen *Auswertebereich* unterteilt. Innerhalb des Integrationsbereichs existieren der *Arbeitsbereich* und die *Basisdatenbank*, innerhalb des Auswertebereichs die *Ableitungsdatenbank*. Der Datenfluss zwischen externen Systemen und dem Data-Warehouse-System ist durch gerichtete Kanten verdeutlicht. Er beginnt bei der Datenquelle und gelangt durch eine Phase der *Extraktion* über den Arbeitsbereich und die Basisdatenbank in die Ableitungsdatenbank. Dabei durchläuft der Datenfluss mehrfach die Phasen der *Transformation* und des *Ladens*. Mit der Phase der *Auswertung* endet der Datenfluss.
Eine Datenquelle existiert ausserhalb des Data-Warehouse-Systems und bezeichnet eine beliebige Quelle, welche die zu integrierenden Daten eines Data-Warehouse-Systems enthält (Bauer & Günzel, 2013). Datenquellen können dabei die Form von strukturierten Daten in Datenbanksystemen oder halbstrukturierten Daten in Dateiform besitzen (Jarke, Lenzerini, Vassiliou & Vassiliadis, 2003).
Die Übertragung der Quelldaten aus der Datenquelle in den Arbeitsbereich wird innerhalb der Extraktionsphase durch eine Extraktionskomponente durchgeführt (Bauer & Günzel, 2013). Dabei existiert eine Extraktionskomponente pro Datenquelle (Bauer & Günzel, 2013). Für die Extraktionsphase kann zur Festlegung von Extraktionszeitpunkten eine Strategie basierend auf Anfrage verwendet werden (Köppen, Saake & Sattler, 2012). Für den Extraktionsvorgang selbst ist es bedeutend, dass dieser auch dann fortgeführt wird, wenn eine Integritätsbedingung nicht erfüllt ist (Bauer & Günzel, 2013; Kimball & Ross, 2013). Ein solches Vorgehen ist dem Umstand geschuldet, dass bereits geladene Daten aus ökonomischen Gründen nicht wiederholt geladen werden

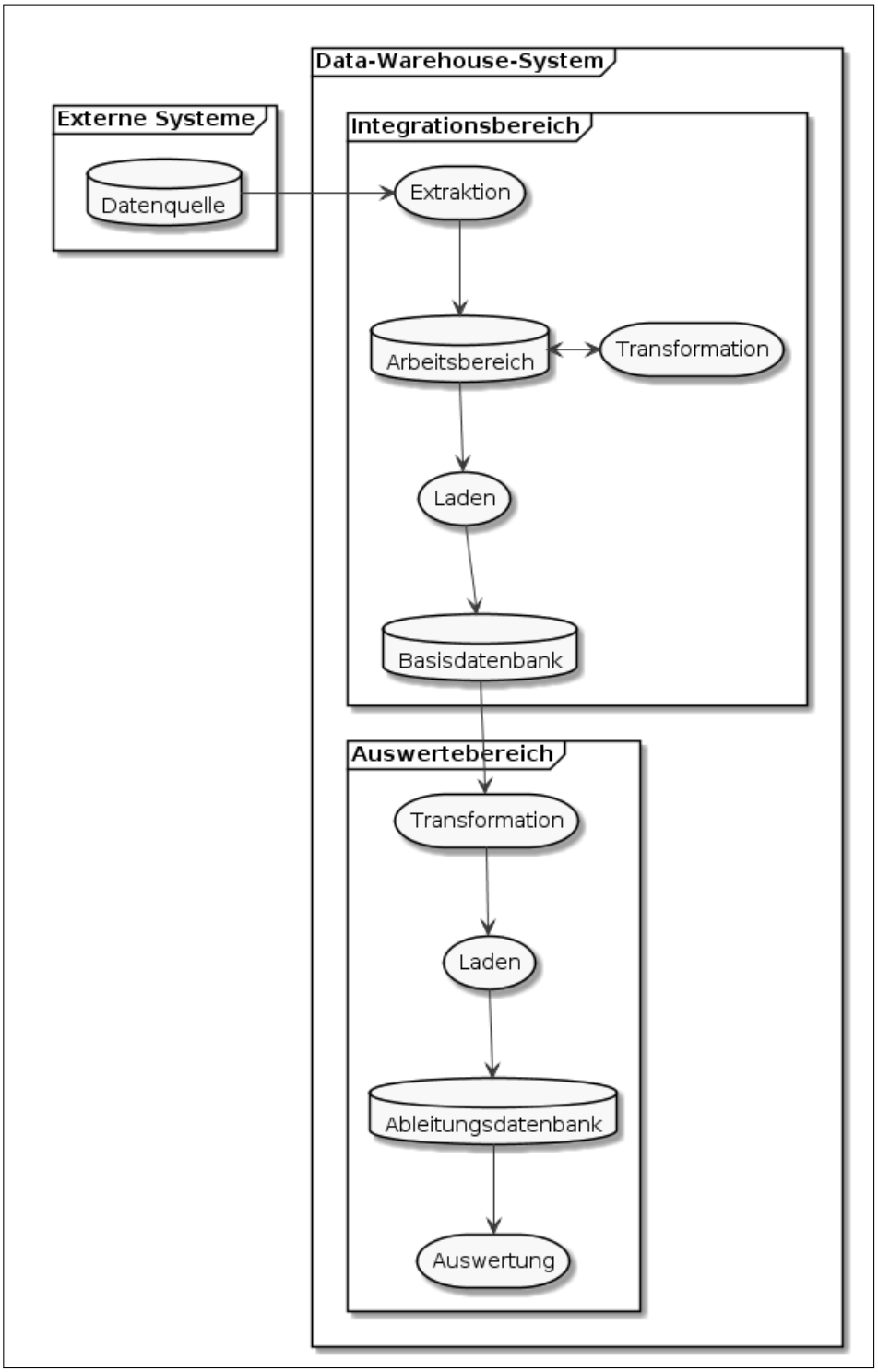

Abbildung 2: Modifizierte Referenzarchitektur eines Data-Warehouse-Systems (Quelle: In Anlehnung an Bauer und Günzel (2013, S. 42))

müssen. Das Auftreten einer nicht erfüllten Integritätsbedingung kann für eine weiterführende Behandlung von dieser geloggt werden. Der Arbeitsbereich wird für das Zwischenspeichern von Daten aus deren Quellen genutzt (Bauer & Günzel, 2013). Er ist von temporärer Struktur und dient hauptsächlich der Transformation von Quelldaten (Köppen et al., 2012). Die Transformation der Daten in ein internes Format des Data-Warehouse-Systems wird durch eine Transformationskomponente vorgenommen (Bauer & Günzel, 2013). Bei der Überführung besteht die Notwendigkeit von Fehlerkorrekturfunktionen zur automatischen Erkennung und Behebung gut definierter Fehler, welche hauptsächlich syntaktischer Art sind (Müller, 2000). Ein solcher Fehler kann beispielsweise ein fehlerhafter Datum-String sein, dessen Delimiter Punkte anstatt Striche aufweist. Während der Ladephase werden analyseunabhängige Detaildaten durch eine Ladekomponente aus dem Arbeitsbereich in die Basisdatenbank geladen (Bauer & Günzel, 2013).
Extraktions-, Transformations- sowie Ladekomponenten – auch als ETL-Komponenten bezeichnet (Bauer & Günzel, 2013) – benötigen nach Behme und Mucksch (2001) für den Zugriff auf die verschiedenen Datenquellen und die Basisdatenbank entsprechende Schnittstellen wie z.B. API-basierte Schnittstellen. Auf Basis von solchen APIs können „sehr problemspezifische Extraktionsroutinen“ (Behme & Mucksch, 2001, S. 44) entwickelt werden.
Die Basisdatenbank besitzt eine integrierte Datenbasis (Bauer & Günzel, 2013). Die vornehmliche Aufgabe der Basisdatenbank ist es, bereinigte Daten zur Verfügung zu stellen (Köppen et al., 2012). Dabei kann nach Behme und Mucksch (2001) eine Basisdatenbank mit Hilfe eines relationalen Datenbanksystems implementiert werden, da ein solches unter anderem einen hohen Standardisierungsgrad aufweist.
Die Phase Transformation im Auswertebereich beschreiben Bauer und Günzel (2013) als Überführung der integrierten Daten aus der Basisdatenbank in ein analyseorientiertes Format. In der anschließenden Ladephase werden den beiden Autoren zufolge die transformierten, analysespezifischen Daten in der Ableitungsdatenbank persistiert, die sie als Grundlage der Auswertung bezeichnen. Denn die Ableitungsdatenbank gewährleistet die Verwaltung und Bereitstellung der für die Analyse benötigten Daten in geeigneter Form. Das

Schema einer Ableitungsdatenbank ist hierfür ausschließlich auf die Benutzerbedürfnisse der Analyse ausgerichtet (Bauer & Günzel, 2013; Köppen et al., 2012). Dabei ist die Normalisierung[4] der Daten innerhalb des Designs nicht so hoch wie bei operativen Datenbanken (Lusti, 2002). Ebenso wie eine Basisdatenbank kann auch eine Ableitungsdatenbank mit Hilfe eines relationalen Datenbanksystems realisiert werden (Bauer & Günzel, 2013; Behme & Mucksch, 2001).
Die abschließende Phase der Auswertung wird von Bauer und Günzel (2013) als Ableitung zweckdienlicher Informationen beschrieben. Diese Ableitung wird mit Hilfe von selbst entwickelten oder durch Hersteller vertriebene Werkzeuge erreicht. Solche Werkzeuge können gemäß der Autoren Tabellen, Grafiken, Texte oder auch multimediale Elemente umfassen.

2.1.2 Data Mining

Beim Data Mining kommen spezielle Algorithmen zum Einsatz, um Muster in Daten zu finden (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Witten, Frank, Hall & Pal, 2017). Dabei können sowohl statistische Methoden als auch Verfahren des ML ohne vorherige Formulierung einer exakten Fragestellung eingesetzt werden (Bauer & Günzel, 2013).

Als de-facto-Standard beim Vorgehen von Data Mining wird von Chapman et al. (2000) der CRISP-DM bezeichnet. Nach ihnen ist dieser branchen-, werkzeug- und auch applikationsneutral und beruht nicht auf theoretischen Prinzipien, sondern auf praktischen Erfahrungen aus Data-Mining-Projekten (Chapman et al., 2000). Der CRISP-DM findet nicht nur in wirtschaflichen Bereichen Anwendung (Chapman et al., 2000), sondern auch im Umfeld von Forschung (Larose, 2005; Bellazzi & Zupan, 2008).
Der Prozess beschreibt den vollständigen Lebenszyklus eines Data-Mining-Projekts (Larose, 2005) und gliedert sich in sechs Phasen, die miteinander in Beziehung stehen (Chapman et al., 2000). Die Phasen des CRISP-DM werden oft wiederholt durchlaufen, bis anwendbare Muster gefunden werden

[4]Siehe auch Faeskorn-Woyke, Bertelsmeier, Riemer und Bauer (2007).

(Larose, 2005).
Der Prozess beginnt, wie aus Abb. 3 auf S. 20 zu entnehmen ist, mit der initialen Phase *Business/ Research Understanding*. Diese ist durch die Ver-

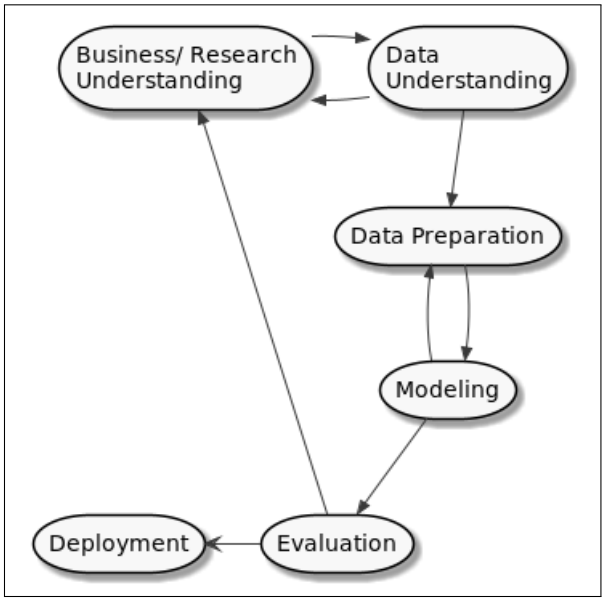

Abbildung 3: CRISP-DM (Quelle: In Anlehnung an Chapman et al. (2000, S. 10) und Larose (2005, S. 6))

ständnisbildung der Projektziele und -anforderungen aus fachlicher Sicht gekennzeichnet (Chapman et al., 2000; Larose, 2005).
Die daran anschließende Phase *Data Understanding* dient der Gewinnung erster Einsichten in die Daten (Chapman et al., 2000). Des Weiteren werden innerhalb dieser Phase auch Qualitätsprobleme identifiziert und beurteilt (Chapman et al., 2000; Larose, 2005). Von der Phase Data Understanding kann auch in die Phase Business/ Research Understanding zurückgekehrt werden, um eventuell notwendige Änderungen an den Projektzielen und -anforderungen vorzunehmen.
In der folgenden Phase *Data Preparation* werden aus den initialen Rohdaten finale Datensätze erstellt (Larose, 2005). Für deren Erstellung findet eine Auswahl von Tabellen, Datenzeilen und Attributen, sowie die Transformation und Bereinigung der Daten statt (Chapman et al., 2000).
Larose (2005) beschreibt diese Phase als sehr arbeitsintensiv, rechtfertigt den

Aufwand jedoch durch die Verwendung der erstellten Datensätze in den folgenden Phasen des CRISP-DM.
In der *Modeling*-Phase findet nach Chapman et al. (2000) die Auswahl und Anwendung diverser Modeling-Techniken sowie deren Parameter-Kalibrierung mit optimalen Werten statt. Abhängig von dem ausgewählten Data-Mining-Verfahren, muss eine andere Aufbereitungsform der Datensätze gewählt werden. In einem solchen Fall ist die Phase Data Preparation wiederholt durchzuführen, um die Datensätze anzupassen.
Die Phase *Evaluation* dient der Evaluierung des Modells und dem Review der bis hier durchgeführten Phasen, wie die Autoren schreiben. Des Weiteren wird gemäß der Autoren in der aktuellen Phase die Erfüllung der Projektziele überprüft und sichergestellt.
Die *Deployment*-Phase ist die letzte Phase im CRISP-DM. Diese dient laut der Autoren dazu, das erlangte Wissen geeignet darzustellen und dem Endnutzer zur Verfügung zu stellen. In Abhängigkeit von den Projektanforderungen kann dabei die Wissensvermittlung von einem Report bis hin zu der Implementierung eines wiederholbaren Data-Mining-Prozesses reichen.

2.2 Machine Learning

Dieses Unterkapitel beschreibt das unüberwachte und das überwachte Lernen, die beide Teilgebiete des ML sind. Für beide Teilgebiete wird jeweils ein Verfahren vorgestellt, das in der vorliegenden Arbeit zum Einsatz kommt.

2.2.1 Unüberwachtes Lernen

Unüberwachtes Lernen dient dem Finden von Strukturen in Daten, ohne dabei Bezug auf Klassenattribute nehmen zu können (Raschka, 2017; Liu, 2011). Dabei lässt sich das unüberwachte Lernen bei Daten mit unbekannter Struktur einsetzen (Raschka, 2017). Eine Teildisziplin des unüberwachten Lernens bildet dabei das Clustering (Liu, 2011).

Clustering dient nach Bacher, Pöge und Wenzig (2010) dem Zusammenfassen einer Menge von Objekten wie beispielsweise von Individuen in homogene Gruppen. Das Grundprinzip der Clusterbildung basiert dabei auf dem Begriff der „homogenen“ Gruppe, nach dem innerhalb eines Clusters Homogenität und zwischen den Clustern Heterogenität vorherrschen soll. Diese beiden Kriterien bilden laut der Autoren die wichtigsten Anforderungen für ein Clustering. Nach forschungspraktischen und inhaltlichen Aspekten kommen noch weitere Anforderungen hinzu, so dass Bacher et al. (2010) für ein Clustering insgesamt den folgenden Kriterienkatalog aufstellen:

1. Homogenität innerhalb der Cluster
2. Heterogenität zwischen den Clustern
3. Erklären der Variation innerhalb der Daten
4. Stabilität der Cluster
5. Interpretierbarkeit der Cluster
6. Valide Cluster
7. Geringe Clusteranzahl

8. Mindestgröße der einzelnen Cluster

Die beiden ersten Kriterien wurden bereits als grundlegende Kriterien weiter oben im Text erläutert. Des Weiteren soll mit Hilfe der Cluster die Variation innerhalb der Daten (Kriterium 3) erklärt werden können. Darüber hinaus soll die Stabilität der einzelnen Cluster (Kriterium 4) gewährleistet sein. Dies ist dann gegeben, wenn geringfügige Änderungen am Datenmaterial nicht zu entscheidenden Änderungen der Ergebnisse der Clusteranalyse führen. Es sollte auch eine gute Interpretierbarkeit der Cluster (Kriterium 5) gegeben sein, so dass nach Möglichkeit inhaltlich sinnvolle Namen für die jeweiligen Cluster vergeben werden können. Des Weiteren sollten die Cluster inhaltlich möglichst valide sein (Kriterium 6). Dies kann garantiert werden, wenn eine Korrelation von Clustern mit externen Variablen vorhanden ist. Voraussetzung dafür ist, dass die externen Variablen möglichst in Zusammenhang mit den Clustern stehen, aber nicht in den Clustern enthalten sind. Die Anforderung einer kleinen und überschaubaren Anzahl (Kriterium 7) an Clustern sowie die einer gewissen Mindestgröße von Clustern (Kriterium 8) erleichtern zum einen die inhaltliche Interpretierbarkeit (Kriterium 5) und zum anderen erhöhen sie die Stabilität der Cluster (Kriterium 4). Darüber hinaus trägt eine gewisse Mindestgröße der einzelnen Cluster laut Aussage der Autoren ebenfalls zu stabilen Clustern (Kriterium 4) bei.

Nach einem erfolgten Clustering ist eine Prüfung sowie die Validierung des gefundenen Modells anhand der folgenden Punkte durchzuführen (Bacher et al., 2010).

1. Prüfung der Modellanpassung
2. Prüfung der inhaltlichen Interpretierbarkeit
3. Stabilitätsprüfung
4. Inhaltliche Validitätsprüfung

Die *Prüfung der Modellanpassung* kann anhand der Kriterien 1 bis 3 sowie 7 und 8 für ein Clustering vorgenommen werden.[5] Eine gute Modellanpas-

[5]Siehe Kapitel 2.2.1 ab S. 22.

sung ist die notwendige Voraussetzung für die anschließende *Prüfung der inhaltlichen Interpretierbarkeit* (Punkt 2). Die *Stabititätsprüfung* überprüft Veränderungen der Ergebnisse bei geringfügigen Modifikationen in den Daten (Punkt 3). Eine *inhaltliche Validitätsprüfung* (Punkt 4) kann laut Meinung der Autoren vorgenommen werden, indem mit Hilfe von Merkmalen, die nicht in die Clusterbildung einbezogen werden, Hypothesen über den Zusammenhang der ermittelten Cluster formuliert werden.

Es existieren verschiedene Verfahren zum Auffinden von Clustern. Für diese Arbeit relevant ist die Vorgehensweise der agglomerativen hierarchischen Clusteranalyseverfahren. Das Vorgehen bei einem solchen Verfahren kann dem *Primitiven Clustering Algorithmus* (Algorithmus 1 auf S. 24) nach Müllner (2011) entnommen werden.

Algorithmus 1 Primitiver Clustering Algorithmus (Quelle: In Anlehnung an Müllner (2011, S. 6))

procedure PRIMITIVE_CLUSTERING(S, d) ▷ S: node labels, d: pairwise dissimilarities
  $N \leftarrow |S|$ ▷ Number of input nodes
  $L \leftarrow []$ ▷ Output list
  $size[x] \leftarrow 1 \; for \; all \; x \; \in \; S$
  for $i \leftarrow 0, ..., N-2$ **do**
    $(a, b) \leftarrow argmin_{(SxS)\setminus\Delta} d$
    Append $(a, b, d[a, b])$ to L.
    $S \leftarrow S\backslash a, b$
    Create a new node label $n \; \notin \; S$.
    Update d with the information

$$d[n, x] = d[x, n] = FORMULA(d[a, x], d[b, x], d[a, b], size[a], size[b], size[x])$$

    for all $x \in S$.
    $size[x] \leftarrow size[a] + size[b]$
    $S \in S \cup n$
  end for
  return L ▷ the stepwise dendrogram, an $((N-1) \times 3)$-matrix
end procedure

$$dist(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ir} - x_{jr})^2} \quad (1)$$

1: Euklidische Distanz (Quelle: Liu (2011, S. 152))

$$\sqrt{\frac{(n_I + n_K)d(I,K) + (n_J + n_K)d(J,K) - n_k d(I,J)}{n_I + n_J + n_K}} \quad (2)$$

2: Ward-Verfahren (Quelle: Müllner (2011, S. 7))

Zeile 1 und 16 beinhalten die Definition und das Ende des Algorithmus. Wie Zeile 1 zu entnehmen ist, nimmt der Algorithmus die zu clusternden Objekte S sowie eine paarweise Unähnlichkeitsmatrix d entgegen. Die paarweise Unähnlichkeitsmatrix kann zum Beispiel durch die Euklidische Distanz (Formel 1 auf S. 25) abgebildet werden. Als erstes wird durch den Algorithmus die Anzahl an Objekten in N gespeichert (Zeile 2), anschließend wird die leere Ausgabeliste L (Zeile 3) erstellt. Das Array *size* enthält die Anzahl an Objekten für jedes Cluster. Da bei einem hierarchisch agglomerativen Algorithmus jedes Objekt zunächst ein eigenes Cluster bildet (Raschka, 2017), wird die Größe für jedes Cluster auf 1 gesetzt (Zeile 4). Im Anschluss werden $N - 1$ Iterationen durchgeführt, um die einzelnen Cluster auf der Basis der Unähnlichkeitsmatrix zu vereinen (Zeile 5 bis 14). Dazu werden zunächst die beiden Cluster ermittelt, die die geringste Unähnlichkeit aufweisen (Zeile 6). Danach werden die beiden Cluster zu einem vereint und zu L hinzugefügt (Zeile 7) und aus S entfernt (Zeile 8). Anschließend erhält das neu gebildete Cluster eine Bezeichnung, die noch nicht in S vorhanden ist (Zeile 9). Daran anschließend ist die Unähnlichkeitsmatrix zu aktualisieren, da ein neues Cluster den Clustern (S) hinzugefügt wurde (Zeile 10) (Müllner, 2011). Diese Aktualisierung geschieht nach einem bestimmten Vorgehen, wie beispielsweise dem Ward-Verfahren, und wird für das neue Cluster n sowie alle verbliebenen Cluster $x \in S$ durchgeführt. Das Ward-Verfahren (Formel 2 auf S. 25) ermittelt dabei die Distanz zwischen den Clustern I, J und K. n bezeichnet dabei die Anzahl an Elementen eines Clusters, d die Distanz zwischen zwei Clustern. Zunächst wird die Summe aus den Produkten von

der Anzahl der Elemente und den Distanzen von Cluster I und K sowie J und K gebildet. Anschließend wird von dieser Summe das Produkt von der Anzahl an Elementen von Cluster K und der Distanz von Cluster I und J substrahiert. Das Ergebnis dieser Substraktion wird danach durch die Summe der Anzahl an Elementen von Cluster$I$, J und K dividiert. Abschließend wird die Quadratwurzel gezogen.
Nach der Berechnung der neuen Distanz wird das Array *size* für Cluster n aktualisiert, indem die Größen der beiden ursprünglichen Cluster a und b addiert werden. Am Ende eines Iterationsschritts wird das neu gebildete Cluster n zur Menge der übrigen Cluster S hinzugefügt (Zeile 13).
Der Algorithmus endet mit der Rückgabe der Ausgabeliste L, welche das Dendogramm mit den einzelnen Schritten der Clusterzusammenführungen enthält (Zeile 15).
Aufgrund der Tatsache, dass bei manchen Clustering-Algorithmen nur die Merkmale und die Objekte spezifiziert werden müssen und dass auch die Anzahl der Cluster offen gelassen werden kann, eignen sich solche Algorithmen für die explorative Datenanalyse (Bacher et al., 2010).

Kritik üben Bacher et al. (2010) beim Einsatz von Clusterverfahren in Publikationen, bei denen keine genaue Spezifizierung der eingesetzten Verfahren beschrieben wird. Die Autoren nennen die folgenden Kriterien für eine solche Spezifizierung:

- Ausgewählte Variablen
- Ausgewählte Objekte
- Ggf. eingesetzte Datentransformationen
- Gewähltes (Un-)Ähnlichkeitsmaß
- Ausgewähltes Verfahren
- Eingesetztes Computerprogramm
- Technische Voreinstellungen

- Verwendete Kriterien zur Bestimmung der Clusterzahl
- Durchgeführte Stabilitätsprüfung
- Durchgeführte Validitätsprüfung

Durch eine ungenaue Spezifizierung wird nach Aussage der Autoren eine Einschätzung und Beurteilung des durchgeführten methodischen Vorgehens vereitelt.

2.2.2 Überwachtes Lernen

Überwachtes Lernen gewinnt nach Liu (2011) Wissen aus Daten, indem dieses durch ein Classification Model abgebildet wird. Die Vorgehensweise zur Modellerstellung wird dabei in Trainings- und Testphase unterteilt. Innerhalb der Trainingsphase wird das Modell mit Hilfe von Trainingsdaten erstellt. Im Anschluss daran wird das Modell unter Verwendung der Testdaten überprüft. Trainings- und Testdaten werden aus den zur Verfügung stehenden Datensätzen gebildet. Dabei dürfen laut Autor die Testdaten nicht bereits zur Erstellung des Modells verwendet werden. Aufgrund des erstellten Modells lassen sich im Anschluss Voraussagen über unbekannte Daten treffen (Raschka, 2017).

Im Folgenden wird mit Hilfe der Pseudocodebasis des rekursiven Algorithmus *decisionTree* nach Liu (2011) (Algorithmus 2 auf S. 28) die Funktionsweise eines Algorithmus für das Erstellen eines Decision Trees erläutert. Der hier dargestellte Algorithmus partitioniert die an ihn übergebenen Daten auf der Basis der Strategie divide-and-conquer.
Der Algorithmus erhält als Startparameter die Trainingsdatensätze (D), die Menge an Attributen (A) und den Decison Tree (T). Der Startpunkt des Algorithmus mit allen Trainingsdatensätzen ist an dessen Wurzel. Durch die rekursive Unterteilung der Datensätze wächst der erzeugte Decision Tree. Es existieren zwei Abbruchbedingungen für die Rekursion. Diese sind in Zeile 2 und 4 zu finden. Nach der ersten Bedingung in Zeile 2 wird T zu einem Blatt

Algorithmus 2 decisionTree Algorithmus (Quelle: In Anlehnung an Liu (2011, S. 70))

```
1: procedure DECISIONTREE(D, A, T)
2:     if D contains only training examples of the same class c_j ∈ C then
3:         make T a leaf node labeled with class c_j;
4:     else if A = ∅ then
5:         make T a leaf node labeled with c_j, which is the most frequent
           class in D;
6:     else ▷ D contains examples belonging to a mixture of classes. We
           select a single attribute to partition D into subsets so that
           each subset is purer
7:         p_0 = impurityEval-1(D);
8:         for each attribute A_i ∈ A(= {A_1, A_2, ..., A_k}) do
9:             p_i = impurityEval-2(A_i,D);
10:        end for
11:        Select A_g ∈ {A_1, A_2, ..., A_k} that gives the biggest impurity
           reduction, computed using p_0 − p_i;
12:        if p_0 − p_g < threshold then ▷ A_g does not significantly reduce
                                          impurity p_0
13:            Make T a leaf node labeled with c_j, the most frequent class
               in D;
14:        else ▷ A_g is able to reduce impurity p_0
15:            Make T a decision node on A_g;
16:            Let the possible values of A_g be v_1, v_2, ..., v_m. Partition D into
               m disjoint subsets D_1, D_2, ..., D_m based on the m values of A_g.
17:            for each D_j in {D_1, D_2, ..., D_m} do
18:                if D_j ≠ ∅ then
19:                    create a branch (edge) node T_j for v_j as a child node
                       of T;
20:                    decisionTree(D_j, A − {A_g}, T_j) ▷ A_g is removed
21:                end if
22:            end for
23:        end if
24:    end if
25: end procedure
```

gemacht und mit dem Klassennamen c_j versehen (Zeile 3), wenn D nur Datensätze der gleichen Klasse enthält. Die zweite Bedingung in Zeile 4 fordert, dass keine Attribute vorhanden sind. In diesem Fall wird T zu einem Blattknoten gemacht und mit der Klasse c_j versehen, welche die häufigste Klasse in D ist (Zeile 5). Sind die beiden Abbruchbedingungen für die Rekursion nicht erfüllt, so werden die Schritte ab Zeile 6 beziehungsweise Zeile 7 des Algorithmus ausgeführt. Dabei wird zunächst die Unreinheit p_0 für D mit Hilfe der Funktion *impurityEval-1* bestimmt (Zeile 7). Die Bestimmung erfolgt anhand der Entropie basierend auf der Informationstheorie nach Shannon (1948) wie die folgende Formel (3) zeigt.

$$entropy(D) = -\sum_{j=1}^{|C|} Pr(c_j) log_2 Pr(c_j), \quad \sum_{j=1}^{|C|} Pr(c_j) = 1 \tag{3}$$

3: Entropie (Quelle: Liu (2011, S. 72))

Die Gleichung bildet die negierte Summe der Produkte der Auftrittswahrscheinlichkeit ($Pr()$) der Klasse c_j in D und dem Logarithmus mit der Basis 2 der jeweiligen Auftrittswahrscheinlichkeit[6]. Für den Logarithmus gilt dabei die Definition $0log(0) = 0$. Je geringer der berechnete Wert für die Entropie ausfällt, desto höher ist die Datenreinheit in D (Liu, 2011).
Nach Bestimmung der Unreinheit wird innerhalb der Iteration durch den Aufruf der Funktion *impurityEval2* für jedes Attribut A_i überprüft, wie stark das jeweilige Attribut die Unreinheit in D reduziert (Zeile 8-10). Diese Funktion basiert wiederum auf der Funktion *gain* aus der folgenden Formel (4).

$$gain(D, A_i) = entropy(D) - entropy_{A_i}(D) \tag{4}$$

4: Gain-Funktion (Quelle: Liu (2011, S. 73))

[6]Die jeweilige Auftrittswahrscheinlichkeit berechnet sich durch den Quotienten der Anzahl aller Beispiele von Klasse c_j in D und allen Beispielen bezwiehungsweise Datensätzen in D, wie Liu (2011) ausführt.

Nach Liu (2011) berechnet die gain-Funktion den Informationsanstieg anhand der jeweiligen Differenz der Entropie von D und der Entropie von D nach der Partitionierung mit Attribut A_i. Die Entropie für den übergebenen Datensatz D wird durch den Formelteil $entropy(D)$ berechnet. Der Formelteil $entropy_{A_i}(D)$ berechnet die Entropie für D, wenn D durch das übergebene Attribut A_i partitioniert wird. Die Entropieberechung für das Attribut A_i ist der folgenden Formel (5) zu entnehmen.

$$entropy_{A_i}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * entropy(D_j) \tag{5}$$

5: Entropie (Quelle: Liu (2011, S. 73))

Innerhalb der Entropie-Berechnung für D mit Attribut A_i stellt v dabei die Anzahl an möglichen Werten für Attribut A_i dar. D wird bei der Verwendung von A_i zur Partitionierung in v disjunkte Datensätze $(D_1, D_2, \ldots, D_v)$ unterteilt.
Nach Durchlaufen der Iteration wird Attribut A_g anhand der Berechnung $p_0 - p_i$ ausgewählt. Dieses bietet die größte Reduktion an Unreinheit für D (Zeile 11). Mit Hilfe der Überprüfung $p_0 - p_i < threshold$ (Zeile 12) wird festgestellt, ob das Attribut A_g die Unreinheit p_0 der Daten im Datensatz signifikant reduziert. Tut es dies nicht, so wird der Baum T zu einem Knotenblatt mit der Bezeichnung c_j, der am häufigsten vorkommenden Klasse in D, gemacht und der Algorithmus stoppt für den Teilbaum (Zeile 13). Ist Attribut A_g jedoch in der Lage, die Unreinheit von p_0 zu reduzieren, so wird der Baum T zu einem Entscheidungsknoten an Attribut A_g gemacht (Zeile 15). Im Anschluss daran werden die m Ausprägungen $v_1, v_2, \ldots, v_m$ des Attributs A_g in die m einzelnen Subdatensätze $D_1, D_2, \ldots, D_m$ geteilt, um den Baum zu erweitern (Zeile 16). Für jeden Subdatensatz (Zeile 17-22), der nicht leer sein darf – wie die Bedingung $D_j \neq \emptyset$ (Zeile 18) fordert – wird anschließend der Teilbaum T_j als Kindknoten des Baumes T für den Wert v_j erstellt (Zeile 19). Anschließend wird der Algorithmus wie Liu (2011) beschreibt mit dem Subdatensatz D_j, den verbleibenden Attributen $A - \{A_g\}$ und dem Teilbaum T_j rekursiv aufgerufen (Zeile 20). Damit endet der Algorithmus.

Mit Hilfe des Algorithmus decisionTree können jedoch nicht nur diskrete, sondern auch kontinuierliche Attribute behandelt werden, wie Liu (2011) angibt. So kann durch binary split der Wertebereich eines Attributs in zwei Intervalle unterteilt werden. Die Unterteilung erfolgt an der Schwelle zwischen den Werten v_i und v_{i+1} in das Intervall $A_i \leq v_i$ und das Intervall $A_i > v_i$, wobei $v_1, v_2, \ldots, v_r$ in aufsteigender Reihenfolge sortiert sind. Jedes Intervall wird dabei als diskretes Attribut betrachtet. Die Berechnung der gain-Formel basiert bei der Behandlung von kontinuierlichen Attributen auf dem jeweiligen Intervall. Es wird nach der Prüfung aller Schwellen die Schwelle ausgewählt, die den Informationsanstieg bestmöglich maximiert, wobei $r - 1$ Schwellen möglich sind. Für die Behandlung von kontinuierlichen Attributen ist Zeile 8-10 anzupassen. Eine weitere Modifikation ist in Zeile 20 vorzunehmen. Das kontinuierliche Attribut A_g darf für kontinuierliche Attribute nicht aus der Rekursion entfernt werden, da es in verschiedenen Teilbäumen mehrfach auftreten kann, wie Liu (2011) angibt.

Aufgrund der Tatsache, dass der Algorithmus decisionTree den Trainingsdatensatz so lange partitioniert, bis keine Unreinheit oder kein Attribut übrig bleiben, besteht die Gefahr des *Overfitting*, wie Liu (2011) schreibt. Bei Overfitting entstehen Bäume mit sehr hoher Tiefe und sehr vielen Blättern. Manche der Teilbäume decken jedoch nur wenige Trainingsdatensätze ab. Die Bäume generalisieren die Daten dann nicht gut, da sie eine hohe Accuracy (Genauigkeit) für die Daten des Trainingsdatensatzes besitzen, jedoch evtl. nicht für die Daten des Testdatensatzes. Overfitting kann durch Rauschen in den Daten, durch falsche Klassenlabel sowie durch falsche Werte der Attribute oder durch die Komplexität und Zufälligkeit des zu untersuchenden Fachgebiets begründet sein.
Overfitting kann nach Angabe des Autors durch *Pruning* (Generalisierung) reduziert werden. Beim Pruning werden Zweige und Teilbäume gelöscht und durch Blätter der Hauptklassen ersetzt. Es existieren verschiedene Methoden für Pruning. Beim Prepruning wird während der Generierung des Baums gestoppt, um Overfitting präventiv zu begegnen. Das effektivere Postpruning entfernt einen Teilbaum nach der vollständigen Generierung des Baums, wenn

der geschätzte Fehler eines Knotens geringer als der geschätzte Fehler des erweiternden Teilbaums ist, so der Autor.

Für die Evaluierung eines Decision Trees wird laut Liu (2011) als Hauptkriterium die *Classification Accuracy* (Fehlerfreiheit) nach der folgenden Formel (6) bestimmt.

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ test\ cases} \tag{6}$$

6: Accuracy-Berechnung (Quelle: Liu (2011, S. 65))

Diese beinhaltet den Quotienten aus der Anzahl an korrekt klassifizierten Testdatensätzen (*Number of correct classifications*) und der Gesamtanzahl an Testdatensätzen (*Total number of test cases*). Je höher die Accuracy, desto genauer ist der Decision Tree.
Des Weiteren kann die Fehlerrate (*Error Rate*) angegeben werden (Liu, 2011). Diese ist nach der folgenden selbsterklärenden Formel (7) zu berechnen.

$$ErrorRate = 1 - Accuracy \tag{7}$$

7: Error Rate (Quelle: Liu (2011, S. 79))

2.3 Ausdauer- und Laboratoriumsdiagnostik

Das vorliegende Unterkapitel gibt zunächst eine Auskunft über die Erhebung der ausdauer- und laboratoriumsdiagnostischen Daten, die in der vorliegenden Arbeit bei der Datenanalyse Verwendung finden. Anschließend werden die der Datenanalyse zugrunde liegenden Parameter aus sportwissenschaflicher Sicht kurz erläutert.

2.3.1 Datenerhebung

Engelmeyer (2012) beschreibt die Erhebung der ausdauer- und laboratoriumsdiagnostischen Daten als Teil der Untersuchungen, die beim Basischeck innerhalb des Deutschen Zentrums für Leistungssport Köln (momentum), erhoben werden. Der Basischeck beinhaltet eine Untersuchung von Athletinnen und Athleten durch momentum, die der Erfassung von deren Gesundheits- und Leistungsstatus dient. Gemäß der Autorin ist der Basischeck ein standardisiertes Messinstrument und in mehrfacher Weise dienlich. Er ermöglicht die Erfassung des vollständigen Athletenstatus an einem Untersuchungstermin. Des Weiteren sind die Analyse-Ergebnisse der erhobenen Daten bei einer Optimierung der Betreuung, der Konkurrenzfähigkeit sowie der Gesunderhaltung und Leistungsverbesserung der Athletinnen und Athleten hilfreich. Zudem soll durch die Analyse das Wissen über die Zusammenhänge zwischen diversen Attributen aus unterschiedlichen Untersuchungsbereichen erweitert werden, so die Autorin.
Die Vielzahl an unterschiedlichen Untersuchungsbereichen im Basischeck wird nach Engelmeyer (2012) dadurch ermöglicht, dass verschiedene Institutionen – als Bestandteile von momentum – an der Durchführung des Basischecks beteiligt sind. Die im Basischeck durchgeführten Untersuchungen können dabei in die Kategorien leistungsdiagnostische Untersuchungen (LDUs) und sportmedizinische Untersuchungen (SMUs) unterteilt werden (Engelmeyer, 2012). Die LDUs beinhalten verschiedene Untersuchungstypen, so auch die Ausdauerdiagnostik.
Um die Ausdauerleistungsfähigkeit zu bemessen, ist nach Engelmeyer (2012) von den einzelnen Individuen ein Stufentest nach einem Messprotokoll in

Anlehnung an Mader (Mader, Liesen & Heck, 1976) zu absolvieren. Die Eingangsgeschwindigkeit bei diesem Test beträgt 2,4 m/s und wird mit jeder Stufe um 0,4 m/s erhöht. Die Steigung bleibt mit einem Prozent konstant. Eine Stufe umfasst dabei eine Dauer von 5 Minuten. Der Belastungsabbruch erfolgt durch das Individuum selbst aufgrund der subjektiven Ausbelastung. Für die Bemessung des Herz-Kreislauf-Systems sowie des Energiestoffwechsels werden laut Engelmeyer (2012) eine Atemgasanalyse, eine Blutkonzentrationsmessung sowie eine Herzfrequenzmessung vorgenommen, wie im Folgenden beschrieben. Der Stufentest selbst wird auf einem Laufband (Woodway ELG 90, Riehl am Rhein) durchgeführt. Mit Hilfe des Analysators ZAN 600 (ZAN MESSGERÄTE GmbH, Oberthulba) findet die Analyse der Atemgase statt. Für die Messung des Laktats – 2- und 4-mmol Schwelle – werden zu Beginn und nach Beendigung einer jeden Stufe sowie 3 Minuten nach Belastungsabbruch 20 μl Blut aus hyperämisierten Ohrläppchen mit Einmal-Glaspipetten entnommen. Analysiert werden die Blutproben mit dem System EBIO PLUS (EPPENDORF, Hamburg). Die Überprüfung der Herzfrequenz wird nach Angaben der Autorin mit einem Brustgurt (Polar Wear Link^{TM}) sowie einer Armbanduhr (Typ S810i, Polar Electro, Kempele, Finnland) durchgeführt.

Die SMUs dienen der Erfassung des Gesundheitsstatus einer Athletin beziehungsweise eines Athleten. Für diese Erfassung kommen auch wiederum verschiedene Untersuchungstypen mit unterschiedlichen Diagnostiken zur Anwendung, so auch die Laboratoriumsdiagnostik. Bei dieser werden nach Engelmeyer (2012) unter anderem hämatologische Parameter des Blutbildes bestimmt und analysiert. Laut der Autorin erfolgt die Blutabnahme für die hämatologischen Parameter durch das Vacutainer System (Becton Dickinson Labware, Heidelberg) bei den einzelnen Individuen nüchtern, im Sitzen und durch temporäres Stauen der Kubitalvene, wobei das Ergebnis 3 ml entnommenes EDTA-Blut (Ethylendiamintetraessigsäure) beträgt. Die Analyse der hämatologischen Parameter findet durch das System Sysmex KX-21N (Sysmex, Norderstedt) statt.

Die Datensätze der beschriebenen Ausdauer- und Laboratoriumsdiagnostik liegen als comma-separated values (CSV)-Dateien vor. Die CSV-Dateien

sind modifizierte Exporte und entstammen dem System eAkte[7] (Engelmeyer, 2012; Nöll, 2009). Neben den verschiedenen Datenspalten weisen die beiden CSV-Dateien jeweils eine Spalte mit einem Stammdaten-Identifier (SD-ID) sowie einem Untersuchungsdatum auf. Durch diese beiden Spalten kann ein Datensatz eindeutig der Untersuchung eines Individuums an einem bestimmten Untersuchungstag zugeordnet werden. Innerhalb der Stammdaten befinden sich auch weitere Informationen zu einem Individuum, die ebenfalls als CSV-Export vorliegen. Die Stammdaten enthalten neben der SD-ID das Geschlecht, das Geburtsdatum sowie die ausgeübte Sportart eines Individuums. Namen oder sonstige Daten, die ein Individuum eindeutig identifizieren, sind in dem Export der Stammdaten und somit auch im Rahmen dieser Arbeit nicht vorhanden.

2.3.2 Parameter

Innerhalb dieser Arbeit wird aus den Attributen der Ausdauer- sowie der Laboratoriumsdiagnostik nur eine Selektion an Parametern bei der Datenanalyse berücksichtigt. Diese Parameter sind in Tabelle 1 auf S. 35 aufgelistet und werden im weiteren Verlauf dieser Arbeit – unterteilt in Leistungsparameter und physiologische Parameter – betrachtet. Zu den Leistungsparametern

Parameter	Abkürzung	Einheit
Zeit bis zum Abbruch	tlim	min
Im Verlauf: Anaerobe Schwelle V4	V4	m/s
Relative maximale Sauerstoffaufnahme (peak)	$rVO_{2_{peak}}$	ml/kg/min
Respiratorischer Quotient (peak)	RQ_{peak}	-
maximale Herzfrequenz	Hf_{max}	S/min
Blutlaktatkonzentration (peak)	Lak_{peak}	mmol/l
Hämoglobin-Wert	Hb	g/dl

Tabelle 1: Parameter der Ausdauer- und Laboratoriumsdiagnostik (Quelle: Eigene Darstellung)

[7] Bei diesem System handelt es sich um das urspüngliche datenhaltende System, welches bei momentum in der Vergangenheit verwendet wurde. Es ist aufgrund der hohen Komplexität bei Administration und Entwicklung sowie der hohen Lizenzkosten nicht mehr im Einsatz.

zählen die Parameter *Zeit bis zum Abbruch* in min und *Im Verlauf: Anaerobe Schwelle V4* in m/s. Die physiologischen Parameter sind die Parameter *relative maximale Sauerstoffaufnahme (peak)* in ml/kg/min, der *Respiratorische Quotient*, die *maximale Herzfrequenz* in S/min und die *Blutlaktatkonzentration (peak)* in mmol/l. Bei den zuletzt genannten Parametern handelt es sich um die Höchstwerte, die beim Abbruch des Stufentests auftreten und bei denen kein leveling off erfolgt (Tomasits & Haber, 2016). Als ein weiterer physiologischer Parameter wird der *Hämoglobin-Wert (Hb)* in g/dl betrachtet, der innerhalb der Laboratoriumsdiagnostik erhoben wird.

Die Zeit bis zum Abbruch (tlim) bezeichnet innerhalb der Spiroergometrie die Summe der gemessenen Zeiten für alle Stufen bis zum symptomlimitierenden Abbruch. Mit der tlim wird angegeben, wie lange ein Individuum die Belastungen des Stufentests aufrecht erhalten kann. Somit stellt die tlim eine Messgröße zur Bestimmung der Ausdauerleistungsfähigkeit dar (Wahl, Bloch & Mester, 2009).
Die Daten der Spiroergometrie innerhalb der vorliegenden Arbeit wurden anhand eines Messprotokolls in Anlehnung an Mader et al. (1976) erhoben.[8] Die Dauer der einzelnen Stufen beträgt dabei 5 min. Da der Abbruch während der Spiroergometrie auch innerhalb einer Stufe möglich war, existieren auch Werte für die tlim, welche ausserhalb von 5 minütigen Stufen liegen (tlim mod $5 \neq 0$).

Die anaerobe Schwelle beschreibt nach Tomasits und Haber (2016) die Belastung, oberhalb derer der Energiebedarf nicht mehr ausschließlich durch die aerobe Energiebereitstellung zu decken ist. Ab dieser Schwelle werden laut der Autoren zusätzlich anaerobe Stoffwechselprozesse zur Deckung des Energiebedarfs hinzugezogen.
Der Übergang von dem maximalen Laktat-steady-state (MLSS) zur anaeroben Energiebereitstellung ist durch diverse Laktatschwellenkonzepte beschrieben (Faude, Kindermann & Meyer, 2009). Für den Kontext dieser Arbeit relevant ist die Laktatschwelle Im Verlauf: Anaerobe Schwelle V4 (V4)

[8] Siehe auch Kapitel 2.3.1 ab S. 33.

nach Mader et al. (1976). Diese ist als ein Bereich beziehungsweise ein Intervall zu verstehen und nicht als eine harte Grenze (Wahl et al., 2009).

Die relative maximale Sauerstoffaufnahme ($rVO_{2_{max}}$) wird von Tomasits und Haber (2016) als Konzentrationsdifferenz von ein- und ausgeatmeter Luft relativ zur Körpermasse beim symptomlimitierenden Abbruch beschrieben.[9] Anhand der $rVO_{2_{max}}$ kann laut der Autoren die Leistungsfähigkeit

$$VO_{2_{max}} = HMV * avDO_2 \tag{8}$$

8: Formel zur Berechnung der $VO_{2_{max}}$ (Quelle: Abgeleitet vom Fick'schen Prinzip $HZV = \frac{V_{O_2}}{(C_a - C_v)_{O_2}}$ [l/min] (Behrends et al., 2012, S. 99))

von Atmung, Kreislauf und Muskelstoffwechsel eines Individuums bemessen werden. Somit besitzt die $rVO_{2_{max}}$ eine wichtige Bedeutung für Ausdauerleistungen im Spitzenbereich (Joyner & Coyle, 2008).
Der höchste Zuwachs für $rVO_{2_{max}}$ kann laut Lundby und Robach (2015) durch High Intensity Intervalltraining erzielt werden. Dabei berufen sich die Autoren auf Hickson, Bomze und Holloszy (1977). Die Trainierbarkeit der $rVO_{2_{max}}$ erscheint jedoch – vor allem aufgrund der Abnahme von der $rVO_{2_{max}}$ mit zunehmendem Alter – limitiert zu sein (Lundby & Robach, 2015). Im Verlauf einer Athletenkarriere kann die $rVO_{2_{max}}$ nach Lundby und Robach (2015) jedoch Werte zwischen 65.2 und 69.1 ml/min/kg bei weiblichen Individuen sowie 77.2 und 83.2 ml/min/kg bei männlichen Individuen annehmen. Dabei können Extremausprägungen bis zu 90.6 ml/min/kg existieren (Burtscher, Nachbauer & Wilber, 2011).
In der vorliegenden Arbeit wird die $rVO_{2_{peak}}$ als Parameter verwendet. Bei dieser findet im Gegensatz zur $rVO_{2_{max}}$ kein leveling off statt (Tomasits & Haber, 2016).

Der RQ (Formel 9, S. 38) bildet das berechnete Verhältnis zwischen Kohlendioxidabgabe ($\dot{V}_{CO_2}$) und Sauerstoffverbrauch ($\dot{V}_{O_2}$) (Silbernagel, 1991).

[9] Innerhalb der vorliegenden Arbeit wird die relative maximale Sauerstoffaufnahme (peak) ($rVO_{2_{peak}}$) in ml/min/kg angegeben.

Nach Kroidl, Schwarz, Lehnigk und Fritsch (2015) ist anhand des RQ erkenn-

$$RQ = \dot{V}CO_2/\dot{V}O_2 \tag{9}$$

9: Formel für den Respiratorischen Quotienten (Quelle: Tomasits und Haber (2016, S. 96))

bar, ob es sich bei der Energiebereitstellung um reine Fettoxidation (RQ im Bereich von 0,7) oder um reine Kohlenhydratverbrennung (RQ im Breich von 1) handelt. Ein Wert $> 1,2$ wird laut der Autoren als Bereich der anaeroben Schuld bezeichnet.
Tomasits und Haber (2016) bezeichnen bereits einen RQ > 1 als anaerobe Schuld. Diese ensteht durch die Freisetzung und Abatmung von CO_2 aus dem Bikarbonatpuffer des Blutes durch Laktat während der laktazid-anaeroben Energiebereitstellung bei hochintensiven Belastungen. Ein RQ-Wert von > 1 gibt nach Angabe der Autoren bereits eine metabolische Ausbelastung eines Individuums wieder.
Innerhalb der vorliegenden Arbeit wird der Respiratorische Quotient (peak) ($\mathrm{RQ_{peak}}$) als Parameter verwendet, der bei Abbruch des durchgeführten Stufentests gemessen wird.

Die Herzfrequenz ist definiert als die Anzahl an Kontraktionen des Herzmuskels pro Minute und kann als lineare Größe zur jeweiligen Belastungshöhe verstanden werden (Tomasits & Haber, 2016). Aufgrund ihres Einflusses auf das Herzminutenvolumen (HMV) ist die Herzfrequenz eine wichtige Größe für die maximale Sauerstoffaufnahme. Sowohl bei Untrainierten als auch bei

$$HMV = SV * HF \tag{10}$$

10: Formel für das HMV (Quelle: Tomasits und Haber (2016, S. 51))

Individuen aus dem Bereich des Hochleistungssports kann die $\mathrm{Hf_{max}}$ nahezu identisch ausfallen (Lundby & Robach, 2015), weshalb diese im Vergleich nicht als ein Maß für sportliche Leistungsfähigkeit betrachtet werden kann.
Die $\mathrm{Hf_{max}}$ wird beim symptomlimitierenden Abbruch der Belastung (Tomasits

& Haber, 2016) in der vorliegenden Arbeit bei Abbruch des Stufentests gemessen.

Nach Wahl et al. (2009) ist die Blutlaktatkonzentration im Blut bedingt durch das Gleichgewicht zwischen Produktion (Glykolyserate) und Elimination von Laktat bei der oxidativen Energiebereitstellung. Dieses Gleichgewicht ist wiederum das Resultat eines multifaktoriellen Prozesses. Des Weiteren heben die Autoren die Bedeutung von Laktat in Prozessen des menschlichen Metabolismus hervor, wie nachfolgend zu lesen ist.
Die Entstehung von Laktat wird von Tomasits und Haber (2016) unter anderem durch eine Kapazitätsüberschreitung der aeroben Energiebereitstellung beschrieben. Zu einer solchen Überschreitung kommt es, wenn das bei der Glykolyse entstehende Pyrovat nicht mehr durch die aerobe Energiebereitstellung der Mitochondrien im Muskel metabolisiert werden kann und dieses dann zu Laktat metabolisiert wird. Ursache für die Umwandlung von Pyrovat in Laktat ist laut der Autoren somit eine für den jeweiligen Belastungsreiz zu geringe Mitochondrienmasse.
Nach seiner Entstehung dient Laktat sowohl als Signalmolekül bei Angio- (Beckert et al., 2006; Trabold et al., 2003) und Vaskulogenese (Milovanova et al., 2008) als auch als Substrat in der oxidativen Energiebereitstellung (Gladden, 2004; Brooks, 2007), wie Wahl et al. anmerken. Darüber hinaus beschreiben sie die positiven Effekte von Laktat auf die Ermüdung eines Skelettmuskels (Karelis, Marcil, Péronnet & Gardiner, 2004). Die Metabolisierung von Laktat in anderen Kompartimenten als dessen Entstehungsort wird durch Transportmechanismen wie Monocarbocylat-Transporter (MCT) ermöglicht, welche dadurch auch wiederum Einfluss auf die Blutlaktatkonzentration besitzen (Bonen, 2000), wie Wahl et al. anführen.
Innerhalb der vorliegenden Arbeit wird für die Messung der Blutlaktatkonzentration der Parameter Blutlaktatkonzentration (peak) ($\mathrm{Lak}_{\mathrm{peak}}$) verwendet. Dieser gibt die Blutlaktatkonzentration bei Belastungsabbruch wieder.

Das Hämoglobin ermöglicht aufgrund seiner Bindungsfähigkeit von Sauerstoff innerhalb der Erythrozyten dessen Transport von den Lungenkapillaren

bis zu den einzelnen Körperzellen (Tomasits & Haber, 2016). Diese Bindung geschieht chemisch aufgrund von 4 Hämgruppen und 4 Globinuntereinheiten (Behrends et al., 2012).
Des Weiteren fungiert Hämoglobin in sehr geringem Maß als Puffer für Säuren, die im Organismus gebildet werden (Tomasits & Haber, 2016).
Bei Olympioniken aus dem Bereich des Ski- sowie des Radfahrens wurden bei männlichen Individuen Werte zwischen 14.1 und 15.7 g/dl und bei weiblichen Individuen zwischen 13 und 13.5 g/dl gemessen (Lundby & Robach, 2015).
Als Referenzwerte geben Bain, Bates, Laffan und Lewis (2012) für männliche Individuen 15 ± 2 g/dl und für weibliche Individuen 13.5 ± 1.5 g/dl an.

3 Material und Methodik

In dem vorliegenden Kapitel wird die Vorgehensweise dieser Arbeit beschrieben und anhand der Grundlagen aus Kapitel 2 ab S. 14 gestützt. In einem nächsten Schritt wird auf die Datengrundlage eingegangen, die für die Datenanalyse innerhalb dieser Arbeit verwendet wurde. Das letzte Unterkapitel führt die verwendete Software sowie Frameworks und Libraries auf.

3.1 Vorgehensweise

Die Vorgehensweise dieser Arbeit basiert auf dem Data-Warehouse-Prozess[10] mit einer modifizierten Referenzarchitektur[11] sowie dem CRISP-DM[12] als integriertem Prozess für die Phase der Analyse und kann den Abbildungen 4 bis 6 von S. 42 bis 43 entnommen werden.[13]
Abb. 4 beinhaltet den ersten Konzeptteil für die Vorgehensweise mit den Datenquellen *Stammdaten*, *Ausdauerdiagnostik* und *Laboratoriumsdiagnostik*[14] sowie das *Data-Warehouse-System* mit dem *Integrationsbereich*. Der Integrationsbereich verfügt über ein *Integrationswerkzeug* für Daten aus CSV-Dateien. Das Werkzeug führt die *Extraktion*, die *Transformation* und das *Laden* durch, um die Stammdaten sowie die Daten der Ausdauer- und Laboratoriumsdiagnostik aus den vorliegenden Quellen in der *Basisdatenbank* zu persistieren.
Der zweite Konzeptteil ist Abb. 5 auf S. 42 zu entnehmen. Die Abbildung enthält das Data-Warehouse-System mit dem Integrations- und *Auswertebereich*. Der Integrationsbereich enthält die Basisdatenbank. Der Auswertebereich weist wiederum ein *Integrationswerkzeug* für CSV-Dateien sowie die *Ableitungsdatenbank* auf. Das Integrationswerkzeug führt das Laden der Daten in die Ableitungsdatenbank aus. Transformationen zur Vorbereitung

[10] Siehe auch Kapitel 2.1.1 ab S. 14.
[11] Siehe Kapitel 2.1.1 ab S. 16.
[12] Siehe auch Kapitel 2.1.2 ab S. 19.
[13] Aus Gründen der Übersicht ist das Konzept der Vorgehensweise in drei Abbildungen unterteilt. Innerhalb der einzelnen Abbildungen sind nur die Bereiche, Prozesse und Komponenten abgebildet, die für den jeweiligen Teil erforderlich sind.
[14] Siehe Kapitel 2.3.1 ab S. 34.

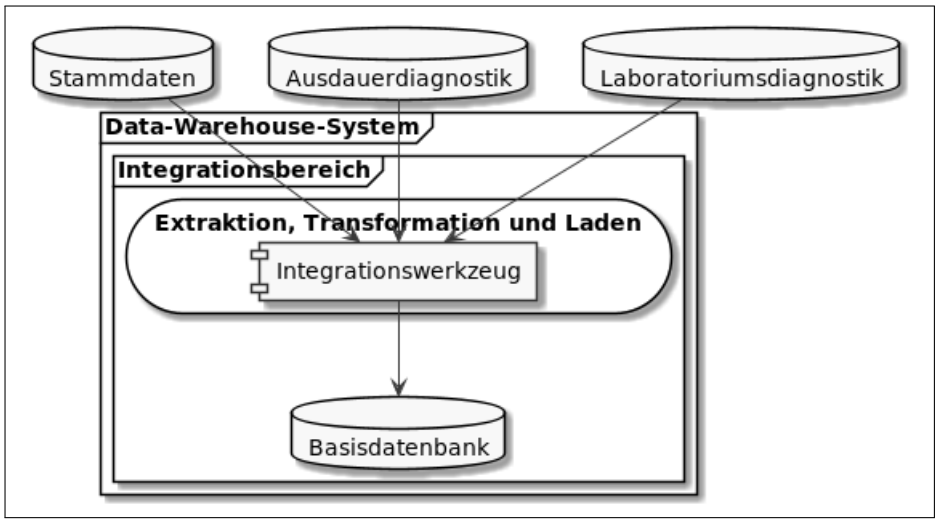

Abbildung 4: Konzeptteil 1 der Vorgehensweise (Quelle: In Anlehnung an Bauer und Günzel (2013, S. 42))

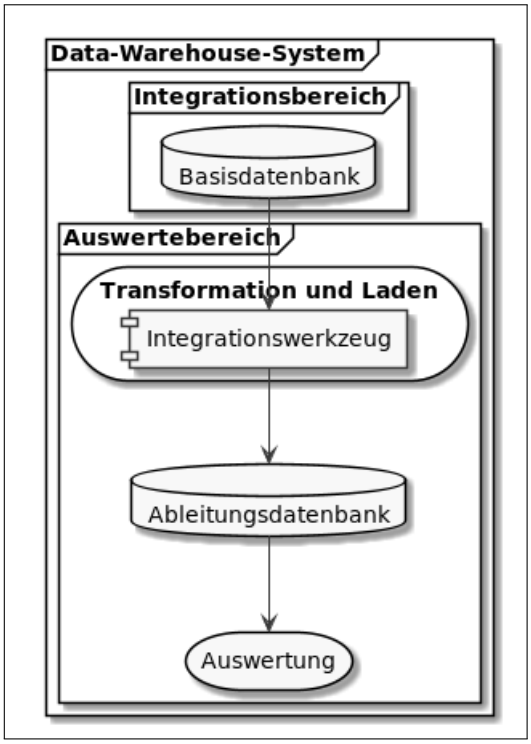

Abbildung 5: Konzeptteil 2 der Vorgehensweise (Quelle: In Anlehnung an Bauer und Günzel (2013, S. 42))

der Analyse wie Aggregationen werden dabei nicht durch das Integrationswerkzeug selbst durchgeführt, sondern durch manuelle Anpassung der Daten.

Im Anschluss an die Integration wird mit Hilfe von CSV-Exporten aus der Ableitungsdatenbank die *Auswertung* der Daten vorgenommen.
Der dritte Konzeptteil für die Vorgehensweise ist Abb. 6 auf S. 43 zu entnehmen. Die Abbildung enthält das Data-Warehouse-System mit dem Auswertebereich. Innerhalb des Auswertebereichs ist die Phase der Auswertung

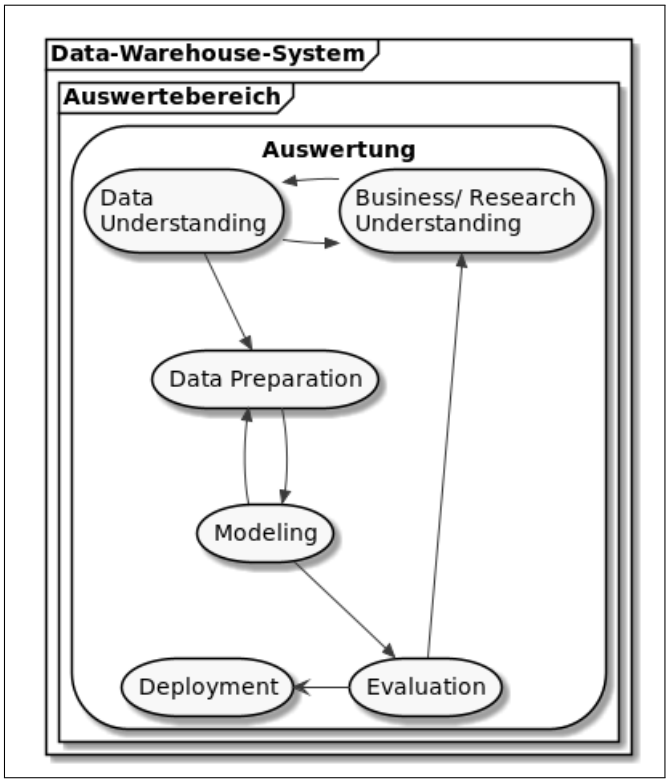

Abbildung 6: Konzeptteil 3 der Vorgehensweise (Quelle: In Anlehnung an Bauer und Günzel (2013, S. 42), Chapman et al. (2000, S. 10) und Larose (2005, S. 6))

dargestellt, welche den CRISP-DM beinhaltet.
Die Phase *Business/ Research Unterstanding* bildet dabei den Ausgangspunkt der Auswertung und ist durch die Zielsetzung[15] dieser Arbeit definiert.

[15]Siehe Kapitel 1.3 ab S. 11.

Im Rahmen dieser Arbeit wurde die Phase *Data Understanding* mit Hilfe von SQL-Selektionen und -Sortierungen innerhalb der Basisdatenbank durchgeführt. Diese wurden auf Datensätze von Stammdaten, Ausdauer- und Laboratoriumsdiagnostik angewendet, um einen Einblick in die Daten zu erhalten. So konnten beispielsweise minimale und maximale Extremwerte identifiziert und auf Plausibilität überprüft werden.
Auch die Phase *Data Preparation* wurde wiederum mit Hilfe der Basisdatenbank durchgeführt. In dieser Phase wurde ein LEFT OUTER JOIN[16] über die Tabellen der Stammdaten, der Ausdauer- sowie der Laboratoriumsdiagnostik angewendet, um eine Verknüpfung zwischen den verschiedenen Datensätzen anhand von SD-ID und Untersuchungsdatum herzustellen. Auf diese Art konnten Datensätze mit den benötigten Parametern V4, $\mathrm{rVO2_{peak}}$, Respiratorischer Quotient (peak) ($\mathrm{RQ_{peak}}$), $\mathrm{Hf_{max}}$, $\mathrm{Lak_{peak}}$ sowie Hb gebildet werden. Diese Datensätze beziehen sich somit auf die Untersuchung eines Individuums an einem Tag. Des Weiteren wurden notwendige Aggregationen für den Parameter tlim manuell durchgeführt. Die so erstellten Datensätze wurden anschließend in der Ableitungsdatenbank gespeichert.
In der Phase *Data Modeling* wurde sowohl auf der Grundlage eines Clusterings[17] als auch auf der eines Decision Trees[18] ein zweistufiges Datenmodell erstellt. Mit Hilfe des Clusterings wurde zunächst basierend auf den Parametern tlim und V4 die Zuordnung eines Individuums zu einem Cluster vorgenommen. Die Durchführung des Clusterings erfolgte mit einem agglomerativen hierarchischen Verfahren sowie dem Ward-Verfahren[19], da zum einen eine geringe Anzahl an Datensätzen vorliegt (Bacher et al., 2010), zum anderen eine gute Nachvollziehbarkeit über die Zusammensetzung der Cluster durch die Visualisierung eines Dendogramms ermöglicht wird. Des Weiteren lieferte das Ward-Verfahren im Vergleich zu anderen Verfahren wie beispielsweise dem Single-Linkage-Verfahren die besten Ergebnisse in Bezug auf den Kriterienka-

[16] Siehe auch Faeskorn-Woyke et al. (2007).
[17] Siehe auch Kapitel 2.2.1 ab S. 22.
[18] Siehe auch Kapitel 2.2.2 ab S. 27
[19] Siehe auch Kapitel 2.2.1 ab S. 25.

talog[20]. Eine Standardisierung[21] der Werte erbrachte für das Clustering keine nennenswerten Veränderungen, so dass auf eine solche bei der Modellbildung verzichtet wurde. Der Grund für die Auswahl der beiden Parameter für das Clustering liegt in dem Umstand begründet, dass mit ihnen eine objektive Aussage über die sportlich erbrachte Leistung getätigt werden kann. Jedes Cluster wurde anschließend als eine sportliche Leistungsklasse betrachtet. Nach Abschluss des Clusterings wurden die Leistungsklassen zur Klassifizierung für die weitere Verarbeitung in dem Klassifikationsverfahren eingesetzt. Aufgrund der guten Nachvollziehbarkeit durch Visualisierung sowie der guten Interpretierbarkeit (Raschka, 2017) wurde als Klassifikationsverfahren ein Decision Tree gewählt. Ein solcher wurde über die Parameter $\mathrm{rVO_2}_{\mathrm{peak}}$, $\mathrm{RQ}_{\mathrm{peak}}$, $\mathrm{Hf}_{\mathrm{max}}$, $\mathrm{Lak}_{\mathrm{peak}}$ sowie Hb erstellt und diente ausschließlich zum Finden von allen Regeln und nicht als Vorhersagemodell. Aus diesem Grund wurde sowohl auf eine Testingphase[22] als auch auf ein Pruning[23] verzichtet.
Die *Evaluation* wurde anhand der Kriterien aus Kapitel 2.2.1 ab S. 22 vorgenommen. Die Stabilitätsanalyse für das Clustering kann Anhang C ab S. 173 entnommen werden. Darüber hinaus wurde die Clusterzugehörigkeit der einzelnen Individuen ergänzend in der Ableitungsdatenbank gespeichert, indem eine neue Tabelle mit den modifizierten Datensätzen angelegt wurde. Auf der Grundlage von sportwissenschaftlicher Literatur wurden die Regeln aus dem Decision Tree durch Selektionen innerhalb der Basisdatenbank auf die modifizierten Datensätze angewendet und konnten so analysiert und entsprechend interpretiert werden.
Als *Deployment* des entstandenen Modells ist das Kapitel 5 ab S. 80 dieser Arbeit aufzufassen.

Sowohl das Data-Warehouse-System als auch das Integrationswerkzeug wurden als Prototypen[24] innerhalb dieser Arbeit implementiert. Des Weiteren

[20] Siehe Kapitel 2.2.1 ab S. 22.
[21] Die Standardisierung überführt eine Zufallsvariable so, dass ihr Mittelwert 0 und ihre Varianz 1 ergibt (Papula, 2001).
[22] Siehe auch Kapitel 2.2.2 ab S. 27.
[23] Siehe auch Kapitel 2.2.2 ab S. 31.
[24] Siehe auch Kapitel 4 ab S. 51.

wurden die Werkzeuge[25] für die Datenauswertung in Form von Skripten implementiert.
Bis zur endgültigen Erstellung des Modells wurde der CRISP-DM in 5 Iterationen durchlaufen.

[25]Siehe auch Kapitel 2.1.1 ab S. 19.

3.2 Datengrundlage

Die in die Betrachtung einfließenden Attribute V4, $rVO2_{peak}$, RQ_{peak}, Hf_{max}, Lak_{peak} und Hb werden direkt aus der Basisdatenbank ausgelesen. Die tlim wird durch die Summe aller absolvierten Zeiten pro Stufe berechnet.

Zur Erstellung des Modells[26] stehen 58 Datensätze mit Individuen aus den Sportarten Boxen, Fußball und Handball zur Verfügung. 27 Individuen sind männlichen und 31 Individuen weiblichen Geschlechts. Die Altersspanne der Individuen liegt zwischen 14 und 25 Jahren.
Die genaue Aufteilung kann der folgenden Tabelle 2 auf S. 47 entnommen werden. Für die Sportart Boxen stehen 15 männliche Individuen für die Modellerstellung zur Verfügung. Fünf Individuen besitzen das Alter von 16, fünf von 17, drei von 18 und jeweils ein Individuum das Alter von 20 und 25 Jahren.

Sportart	Geschlecht	Alter	Anzahl
Boxen	männlich	16	5
		17	5
		18	3
		20	1
		25	1
Fußball	weiblich	14	2
		15	21
		16	6
		17	2
Handball	männlich	15	3
		16	5
		18	4

Tabelle 2: Datengrundlage (Quelle: Eigene Darstellung)

Des Weiteren können 31 weibliche Individuen aus der Sportart Fußball mit in die Erstellung des Modells einbezogen werden. Im Einzelnen existieren zwei Individuen im Alter von 14, einundzwanzig im Alter von 15, sechs im Alter

[26] Siehe Kapitel 3.1 ab S. 44.

von 16 und zwei im Alter von 17 Jahren.
In der Sportart Handball sind 12 männliche Individuen vertreten. Davon besitzen drei das Alter von 15, fünf das Alter von 16 und vier das Alter von 18 Jahren.

3.3 Software

Die vorliegende Arbeit wurde mit Hilfe des Betriebssystems Linux Mint (*Linux Mint (Version 19.1)*, 2018) erstellt.

Der Prototyp der Client/Server Software ist in Java 8 mit Hilfe des OpenJDK (*OpenJDK (Version 11.0.7)*, o. J.) sowie Spring Boot (Webb et al., 2017) realisiert. Die Verbindung zum Datenbankserver wurde mittels MySQL Connector/J (*MySQL Connector/J (Version 8.0.8-dmr)*, o. J.) umgesetzt. Des weiteren wurden die Libraries Apache Commons Lang (*Apache Commons Lang (Version 3.4)*, o. J.), Apache Commons CSV (*Apache Commons CSV (Version 1.5)*, o. J.) und SLF4J (*Simple Logging Facade for Java (SLF4J) (Version 1.7.25)*, o. J.) eingesetzt.
Die Implementierung erfolgte unter Verwendung der IDE IntelliJ IDEA (*IntelliJ IDEA Community (Version 2020.1)*, 2020) sowie des Build-Management-Tools Maven (*Apache Maven (Version 3.6.0)*, 2018). Als relationales Datenbanksystem[27] kommt MySQL Community Server (*MySQL Community Server (Version 5.7.29)*, o. J.) zum Einsatz. Das Testen der API-Endpunkte sowie der Aufruf des Endpunkts zum Erstellen einer Tabelle wurden mit Hilfe der Applikation Postman (*Postman (Version 7.22.1)*, o. J.) durchgeführt.

Die Datenbereinigung sowie die Berechnung der tlim wurden manuell mit Hilfe von LibreOffice Calc (*LibreOffice (Version 6.0.7.3)*, 2018) durchgeführt.
Für die Datenanalyse und die Erstellung der verschiedenen Diagramme – mit Ausnahme der Parallelkoordinatendiagramme – wurden innerhalb dieser Arbeit die Python Data Science Plattform Anaconda (*Anaconda (Version 3.5.1)*, 2018) sowie die IDE PyCharm (*PyCharm Community Edition (Version 2018.1.1)*, o. J.) für Python verwendet. Das Clustering wurde mit der Python Library SciPy (Virtanen et al., 2020) umgesetzt. Dabei wurde die Funktion scipy.cluster.hierarchy.linkage der Library verwendet. Die Abbildung der Dendogramme erfolgte mit der Funktion scipy.cluster.hierarchy.dendrogram der gleichen Library. Der Decision Tree wurde mit Hilfe der Klas-

[27] Siehe auch Kapitel 2.1.1 ab S. 19.

se sklearn.tree.DecisionTreeClassifier aus der Library Scikit-learn (Pedregosa et al., 2011) erstellt, die Abbildung des Decision Trees mit der Funktion sklearn.tree.export_graphviz aus derselben.
Die Erstellung der Boxplots, Balken- und Streudiagramme fand mit Hilfe der Python Library Matplotlib (Hunter, 2007) statt.
Die Parallelkoordinatendiagramme wurden mit Hilfe der Statistiksoftware R (*R (Version 3.4.4)*, o. J.) unter der IDE RStudio (*RStudio (Version 1.2.5019)*, 2019) visualisiert. Dabei kam die Funktion parcoord aus dem Package MASS (Venables & Ripley, 2002) zum Einsatz.

Der Text dieser Arbeit wurde mit Texmaker (Brachet, 2020) sowie LaTeX aus dem Paket TeX Live (*TeX Live (Version 2017.20180305)*, 2018) erstellt. Für die Literaturverwaltung wurde JabRef (*JabRef (Version 3.8.2)*, 2017) verwendet. Die Erstellung der UML Diagramme sowie des Wireframes für das Dashboard wurde auf der Basis von PlantUML (*PlantUML (Version 1.2019.11)*, 2019) mit Hilfe des IntelliJ IDEA Plugins PlantUML Integration (Steinberg, 2020) realisiert.

4 Prototypisches System

Das vorliegende Kapitel gibt einen Überblick über das im Rahmen dieser Arbeit entstandene prototypische System. Dazu werden die Anforderungen an das System sowie die Architektur des Systems beschrieben. Die beiden letzten Unterkapitel geben einen Einblick sowohl in die Umsetzung als auch in die Anwendung des Systems.

4.1 Anforderungen

Die Anforderungen an das im Kontext dieser Arbeit zu entwickelnde System ergeben sich sowohl aus den allgemeinen Anforderungen an ein Data-Warehouse-System[28] als auch aus den spezifischen Anforderungen innerhalb dieser Arbeit. Die Anforderungen an das zu erstellende System sind im Folgenden aufgelistet.

1. **Geringe Lizenzkosten**
   Aufgrund eines eingeschränkten Budgets sind die Lizenzkosten für das System möglichst gering zu halten.

2. **Leichtgewichtiges System**
   Das System ist so zu entwickeln, dass die administrativen Aufwände für den Betrieb des Systems gering gehalten werden können. Aus diesem Grund ist die Entwicklung eines leichtgewichtigen Ansatzes zu verfolgen.

3. **Erweiterung des Datenbankschemas**
   Um weitere Entwicklungskosten einzudämmen, ist ein Mechanismus für die Erweiterung des Datenbankschemas zu verwenden, der ohne Programmierkenntnisse auskommt. Des Weiteren ist zur Beschreibung des Datenbankschemas ein Format zu verwenden, welches für Menschen gut lesbar ist.

[28] Siehe auch Kapitel 2.1.1 ab S. 14

4. **Netzwerkfähigkeit**
   Da nicht gewährleistet werden kann, dass mögliche Datenquellen direkten Zugriff auf das Daten verwaltende System haben werden, ist eine Netzwerkfähigkeit des Systems mit API Voraussetzung.[29]

5. **Authentifizierung**
   Benutzer müssen sich am System anmelden, um dessen Funktionalität nutzen zu können.

6. **Verschlüsselung**
   Die Anbindung und Kommunikation des Systems über das Netzwerk muss aufgrund von Datensicherheit verschlüsselt erfolgen.

7. **Integration von multiplen Datensätzen**
   Das System muss eine Funktionalität zur Verfügung stellen, um mehrere Datensätze eines Typs integrieren zu können. Die Integration muss dabei auch dann fortgeführt werden, wenn Integrationsbedingungen für einen Teil der Datensätze nicht erfüllt sind.[30]

8. **Timestamp**
   Das System muss in der Lage sein, integrierte Datensätze automatisch mit einem Timestamp bei der Persistierung zu versehen.[31]

9. **Logging**
   Das System soll ein Logging über die Integrationstätigkeiten zur Verfügung stellen. Zu diesen zählen sowohl erfolgreiche als auch erfolglose Datenintegrationen.

10. **Integration von CSV-Dateien**
    Das System muss in der Lage sein, Datensätze aus CSV-Dateien integrieren zu können.[32]

[29] Siehe Kapitel 2.1.1 ab S. 18.
[30] Siehe Kapitel 2.1.1 ab S. 16
[31] Siehe Kapitel 2.1.1 ab S. 15.
[32] Siehe Kapitel 3.1 ab S. 41.

11. **Export von CSV-Dateien**
    Es ist eine Exportfunktion für CSV-Datein zur Verfügung zu stellen. Eine solche Funktionalität muss nicht netzwerkfähig sein.[33]

12. **Eindeutigkeit von Spalten**
    Um die Eindeutigkeit eines Datensatzes anhand von Stammdaten-ID[34] und Untersuchungsdatum[35] gewährleisten zu können, ist eine Möglichkeit zu schaffen, einzelne oder mehrere Spalten als eindeutig zu deklarieren.

13. **Aggregation von Datensätzen**
    Das System muss in der Lage sein, Datensätze aus der Ausdauer- sowie der Laboratoriumsdiagnostik miteinander zu aggregieren.[36]

14. **Filter- und Sortierfunktionen**
    Über das System sind Funktionen zur Verfügung zu stellen, die Regelüberprüfungen ermöglichen.[37]

15. **Basis- und Ableitungsdatenbank**
    Das System soll über eine Basisdatenbank[38] für die Datenintegration verfügen sowie über eine Ableitungsdatenbank [39] für Tabellen bezüglich der Auswertung.

[33]Siehe Kapitel 3.1 ab S. 43.
[34]Siehe Kapitel 2.3.1 ab S. 35
[35]Siehe Kapitel 2.3.1 ab S. 35.
[36]Siehe Kapitel 3.1 ab S. 43.
[37]Siehe Kapitel 3.1 ab S. 44.
[38]Siehe Kapitel 3.1 ab S. 41.
[39]Siehe Kapitel 3.1 ab S. 41.

4.2 Architektur

Einen architekturellen Überblick über das im Rahmen dieser Arbeit entstandene System bietet Abb. 7 auf S. 54. Bei dieser Architektur handelt

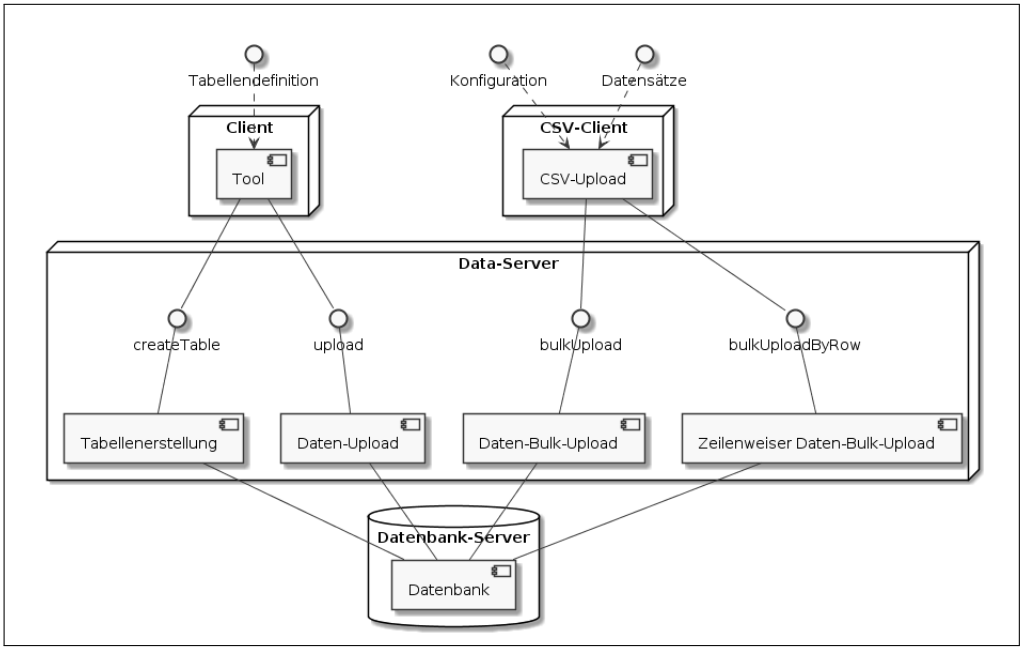

Abbildung 7: Komponentendiagramm *Gesamtüberblick* (Quelle: Eigene Darstellung)

es sich um eine Client-Server-Architektur[40], deren Kommunikation über eine verschlüsselte[41] JavaScript Object Notation (JSON)-Schnittstelle stattfindet. Die Komponenten des Systems können übergeordnet in *Client*, *CSV-Client*[42], *Data-Server* sowie *Datenbank-Server* unterteilt werden. Des Weiteren können die Schnittstellen und Komponenten gemäß ihrer zu erfüllenden Funktion in zwei Teilsysteme unterteilt werden. Diese dienen zum einen dem Anlegen einer Tabelle und damit der Erweiterung des Datenbankschemas[43], zum anderen der Integration von Daten.

[40] Umsetzung der Anforderung *Netzwerkfähigkeit* aus Kapitel 4.1 ab S. 52.
[41] Umsetzung der Anforderung *Verschlüsselung* aus Kapitel 4.1 ab S. 52
[42] Umsetzung der Anforderung *Integration von CSV-Dateien* aus Kapitel 4.1 ab S. 52
[43] Umsetzung der Anforderung *Erweiterung des Datenbankschemas* aus Kapitel 4.1 ab S. 51

Das Anlegen einer Tabelle geschieht mit Hilfe des Clients beziehungsweise dessen Komponente *Tool*, der Schnittstelle *createTable* sowie der Komponente *Tabellenerstellung* innerhalb des Data-Servers. Um eine Tabelle anzulegen, werden mit Hilfe eines geeigneten Clients[44] die Tabellendefinition als JSON eingelesen und über die Schnittstelle createTable eine Anfrage zum Erstellen einer Datenbanktabelle an den Server (Data-Server) gesendet. Dieser verarbeitet die Anfrage und erstellt entsprechend der Tabellendefinition eine Tabelle in der Komponente *Datenbank* innerhalb des Datenbank-Servers.
Der Vorgang der Datenintegration wird durch unterschiedliche Schnittstellen und Komponenten unterstützt. Dabei ist die Integration eines einzelnen Datensatzes sowie multipler Datensätze möglich.
Für die Integration eines einzelnen Datensatzes wird zunächst durch einen entsprechenden Client ein Datensatz als JSON-Objekt ausgelesen und als Request an die Schnittstelle *upload* des Data-Servers gesendet. Innerhalb des Servers wird das JSON-Objekt des Requests durch die Komponente *Daten-Upload* in ein SQL-Insert-Statement umgewandelt, über das der Datensatz in der Datenbank persistiert wird.
Für die Integration von multiplen Datensätzen stellt die Architektur zwei Möglichkeiten bereit. Voraussetzung für beide bildet der Einsatz des CSV-Clients. Dieser liest eine Konfigurationsdatei (Schnittstelle *Konfiguration*) und die eigentliche CSV-Datei (Schnittstelle *Datensätze*) mit multiplen Datensätzen ein. Die Komponente *CSV-Upload* wandelt die Datensätze der CSV-Datei auf der Basis der Konfigurationsdatei in ein JSON-Objekt um, welches anschließend an den Data-Server gesendet wird. Für das Senden von multiplen Datensätzen existieren die beiden Schnittstellen *bulkUpload* und *bulkUploadByRow* innerhalb des Data-Servers. Beide können durch den CSV-Client angesprochen werden. Findet die Übertragung der Daten über die Schnittstelle bulkUpload statt, verarbeitet die Komponente *Daten-Bulk-Upload* das empfangene JSON-Objekt und transformiert es in ein entsprechendes SQL-Insert-Statement, welches auf der Datenbank des Datenbank-Servers ausgeführt wird. Bei dieser Art der Datenintegration wird bei einem Fehler die gesamte Operation abgebrochen und keine Integration durchge-

[44] Siehe Kapitel 3.3 ab S. 49.

führt. Soll die Integration nicht als atomare Operation ausgeführt werden, so nutzt die Komponente des CSV-Clients die Schnittstelle *bulkUploadByRow*. Bei dieser wird das vom CSV-Client übertragene JSON-Objekt von der Komponente *Zeilenweiser Daten-Bulk-Upload* dahingehend verarbeitet, dass jeder Datensatz einzeln per Insert-Statement in der Datenbank persistiert wird. Durch dieses Vorgehen muss eine Datenintegration nicht abgebrochen werden, wenn eine Teilmenge der Datensätze nicht integriert werden kann.[45]

Es existieren jeweils zwei Serverinstanzen der hier vorgestellten Architektur, welche über unterschiedliche Ports adressiert werden können. Jede Serverinstanz ist wiederum mit einer anderen Datenbankinstanz verbunden. Bei den beiden Datenbankinstanzen handelt es sich um eine Basisdatenbank und eine Ableitungsdatenbank.[46]

[45] Umsetzung der Anforderung *Integration von multiplen Datensätzen* aus Kapitel 4.1 ab S. 52.

[46] Umsetzung der Anforderung *Basis- und Ableitungsdatenbank* aus Kapitel 4.1 ab S. 53

```
1  {
2    "tableName": "name_der_tabelle",
3    "tableColumns": [
4      {
5        "type": "bigint",
6        "unique":"true",
7        "name": "name_der_spalte"
8      }
9    ]
10 }
```

Listing 1: JSON-Format für das Anlegen einer Tabelle (Quelle: Eigene Darstellung)

4.3 Umsetzung

In diesem Unterkapitel wird die Umsetzung bzw. die Implementierung der Architektur vorgestellt und beschrieben. Dabei werden die entwickelten Funktionalitäten mit Aktivitäts- und Klassendiagrammen beschrieben.
Die Anwendung beider Diagrammtypen erhebt nicht den Anspruch auf Vollständigkeit. Vielmehr dienen die Diagramme dazu, einen Überblick über das in dieser Arbeit implementierte System zu gewinnen und ein Verständnis für das System zu schaffen. So werden innerhalb der Aktivitätsdiagramme nur die Hauptaktivitäten ohne explizite Detailtreue aufgeführt, um den Programmfluss innerhalb der Funktionalitäten zu verdeutlichen. Zu Gunsten der besseren Lesbarkeit wird des Weiteren innerhalb der Klassendiagramme auf die Darstellung von Komponenten aus den verwendeten Frameworks[47,48] verzichtet. So werden nur die Kernkomponenten der Funktionalitäten dargestellt und diese auch nicht im Detail beschrieben.

4.3.1 Data-Server

Der Endpoint /createTable dient dem Erstellen einer Tabelle innerhalb einer Datenbank auf der Basis eines JSON-Objekts. Das Format dieses Objekts ist Listing 1 auf S. 57 zu entnehmen. Das JSON-Objekt beginnt und schließt mit einer öffnenden beziehungsweise schließenden geschweiften Klammer (Zeile 1

[47] Siehe Kapitel 3.3 ab S. 49

[48] Umsetzung der Anforderungen *Geringe Lizenzkosten* und *Leichtgewichtiges System* aus Kapitel 4.1 ab S. 51.

und 10). Für den Schlüssel *tableName* (Zeile 2) kann als Wert der Tabellenname – hier *name_der_tabelle* – der zu erstellenden Tabelle angegeben werden. Über das Array mit dem Schlüssel *tableColumns* (Zeile 3 bis 9) können einzelne Spalten angegeben werden. Eine Tabellenspalte wird dabei durch ein JSON-Objekt mit den Schlüsseln *type* (Zeile 5), *name* (Zeile 6) sowie dem optionalen Property *unique* (Zeile 7) konfiguriert. Als Wert für den Pflichtschlüssel type können die Werte int, bigint, float, double, date und varchar gepflegt werden. Für den Schlüssel name ist als Wert die Bezeichnung der Spalte einzutragen. Dieser Schlüssel ist mandatorisch. Über den optionalen Schlüssel unique mit den boolschen Werten true und false kann die Spalte als Teil eines Unique-Contraints[49] ausgezeichnet werden. Alle Spalten, die für den Schlüssel unique den Wert true beinhalten, werden zur Bildung des Contraints mit einbezogen.

Die Aktivität für das Erstellen einer Tabelle ist Abb. 8 auf S. 59 zu entnehmen. Nach Beginn der Aktivität erfolgt als erster Schritt die Entgegennahme des JSON-Objekts aus dem Request. Im anschließenden Schritt wird das JSON-Objekt in Plain Old Java Objects (POJOs) überführt und diese anschließend wiederum in ein Create-Statement. Dieses Statement wird im Anschluss ausgeführt. Nach der Ausführung des Statements erfolgt eine Überprüfung auf Fehler, die während der Ausführung des Statements auftreten können. Treten keine solchen auf, wird der HTTP-Status 201 (Created) zurückgegeben und die Aktivität beendet. Im Fehlerfall werden der HTTP-Status 409 (Conflict) sowie eine Meldung entsprechend des Fehlers zurückgegeben.

Abb. 9 auf S. 60 zeigt das Klassendiagramm des Endpoints /createTable mit den Klassen *TableCreatorController*, *TableCreatorService*, *TableCreationJson*, *TableCreatorDao*, *TableCreationException* aus dem Package *org.dataint.server.tablecreation* sowie der Klasse *SqlStatementCreator* aus dem Package *org.dataint.server.utils*, der Klasse *TableColumn* und dem Enum *ColumnType* aus dem Package *org.dataint.server.model*. Die Klasse TableCreatorController nimmt das im Request-Body übertragene JSON-Objekt entgegen.

[49] Umsetzung der Anforderung *Eindeutigkeit von Spalten* aus Kapitel 4.1 ab S. 53

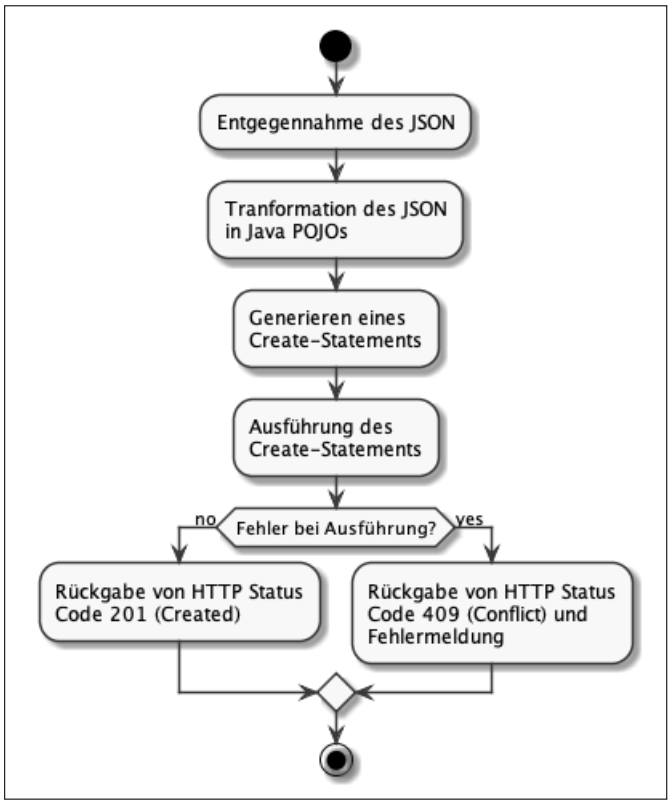

Abbildung 8: Aktivitätsdiagramm *Tabelle erstellen* (Quelle: Eigene Darstellung)

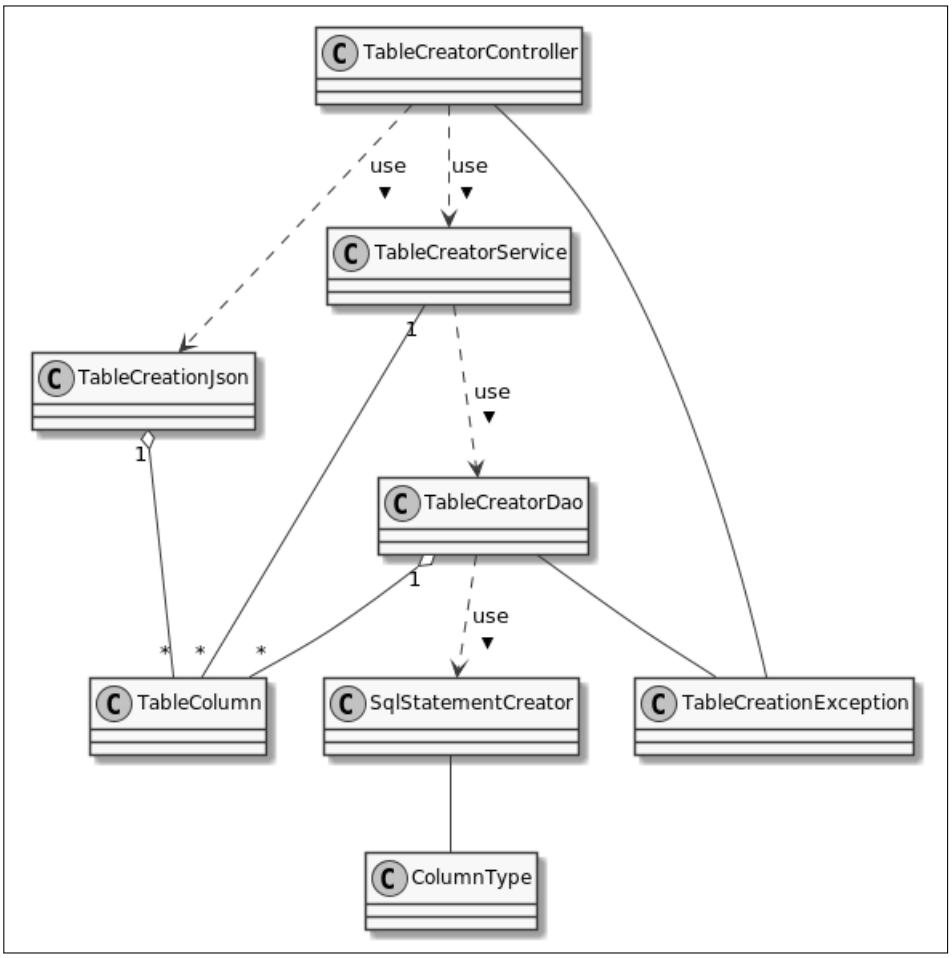

Abbildung 9: Klassendiagramm *Tabelle erstellen* (Quelle: Eigene Darstellung)

Über das verwendete Framework[50] wird das JSON-Objekt auf die Klasse TableCreationJson gemappt. Die Klasse enthält sowohl den Namen der Tabelle als auch eine Liste mit Definitionen von Tabellenspalten. Eine Tabellenspalte wird durch die Klasse TableColumn abgebildet. Eine Instanz der Klasse TableCreationJson besitzt dabei keine, eine oder viele Instanzen der Klasse TableColumn. Der Controller delegiert die Tabellenerstellung an die Klasse TableCreatorService, wobei diese wiederum an die Klasse TableCreatorDao delegiert. Die Klasse SqlStatementCreator generiert unter Zuhilfenahme des Enums ColumnType aus dem übergebenen Datenbanknamen ein Create-Statement, welches durch die SqlStatementCreator-Instanz zusätzlich sowohl eine ID-Spalte mit einem Primärschlüssel als auch eine Spalte mit einem Timestamp[51] erhält. Das Create-Statement wird nach seiner Erstellung von der Klasse TableCreatorDao auf der entsprechenden Datenbank ausgeführt. Führt die TableCreatorDao-Instanz zur Laufzeit ein ungültiges Create-Statement aus, so wirft sie eine Exception vom Typ TableCreationException. Die Klasse TableCreationException ist von der Klasse Exception abgeleitet und wird durch den Aufruf-Stack der Klassen TableCreatorDao und TableCreatorService an die Klasse TableCreatorController durchgereicht. Nach erfolgter Ausführung des Statements gibt die Klasse TableCreatorController die entsprechenden Status-Codes und eventuelle Fehlermeldungen an den Aufrufer zurück.

Für die Integration eines einzelnen Datensatzes ist der Endpoint */dataUpload* zu verwenden. Der Datensatz wird als JSON-Objekt (Listing 2 auf S. 62) an den Endpoint gesendet. Neben dem Namen der Tabelle (Schlüssel *tableName*), in welche der Datensatz zu integrieren ist, werden über das Array *values* JSON-Objekte angegeben, die jeweils einen Wert für eine Tabellenspalte einer Tabellenzeile innerhalb der Datenbank repräsentieren. Ein solches JSON-Objekt enthält sowohl einen Spaltennamen (Schlüssel *columnName*) als auch den entsprechenden Wert (Schlüssel *value*) für die jeweilige Spalte. Das Array muss soviele JSON-Objekte enthalten, wie die Tabelle Spalten besitzt.
Abb. 10 auf S. 63 zeigt die Aktivität *Integration eines einzelnen Datensatzes*.

[50] Siehe auch Kapitel 3.3 ab S. 49.

[51] Umsetzung der Anforderung *Timestamp* aus Kapitel 4.1 ab S. 52.

```
{
  "tableName":"tabellenname",
  "values":
  [
    {"columnName":"name_s_1","value":"'Wert S 1'"},
    {"columnName":"name_s_2","value":"'Wert S 2'"},
    {"columnName":"name_s_3","value":"'Wert S 3'"}
  ]
}
```

Listing 2: JSON-Format für die Integration eines einzelnen Datensatzes (Quelle: Eigene Darstellung)

```
{
  "tableName":"tabellenname",
  "values":
  [
    [
      {"columnName":"name_s_1","value":"'Wert 1 S 1'"},
      {"columnName":"name_s_2","value":"'Wert 1 S 2'"},
      {"columnName":"name_s_3","value":"'Wert 1 S 3'"}
    ]
  ]
}
```

Listing 3: JSON-Format für die Integration multipler Datensätze (Quelle: Eigene Darstellung)

Die Abbildung gibt einen Überblick über den Programmablauf beim Persistieren eines einzelnen Datensatzes innerhalb der Datenbank. Die Aktivität beginnt mit der Transformation des empfangenen JSON-Objekts in POJOs. Anschließend wird aus den transformierten POJOs ein Insert-Statement generiert, welches in einem weiteren Schritt ausgeführt wird. Nach der Ausführung des Statements erfolgt eine Überprüfung auf eine fehlerfreie Ausführung. Liegt eine solche vor, so werden eine Erfolgsmeldung sowie der Status-Code 200 zurückgegeben. Im Fehlerfall werden eine Fehlermeldung und der Status-Code 409 zurückgegeben.
Für die Integration von multiplen Datensätzen stehen die beiden Endpoints /bulkUpload und /bulkUploadByRow zur Verfügung. Diese beiden Endpoints akzeptieren Daten in Form des JSON-Objekts aus Listing 3 auf S. 62. Zeile 1 und 11 enthalten die öffnende und schließende Klammer des JSON-Objekts. In Zeile 2 wird über den Schlüssel *tableName* der Tabellenname festgelegt, der die Tabelle angibt, in welche die Integration erfolgen soll. Von

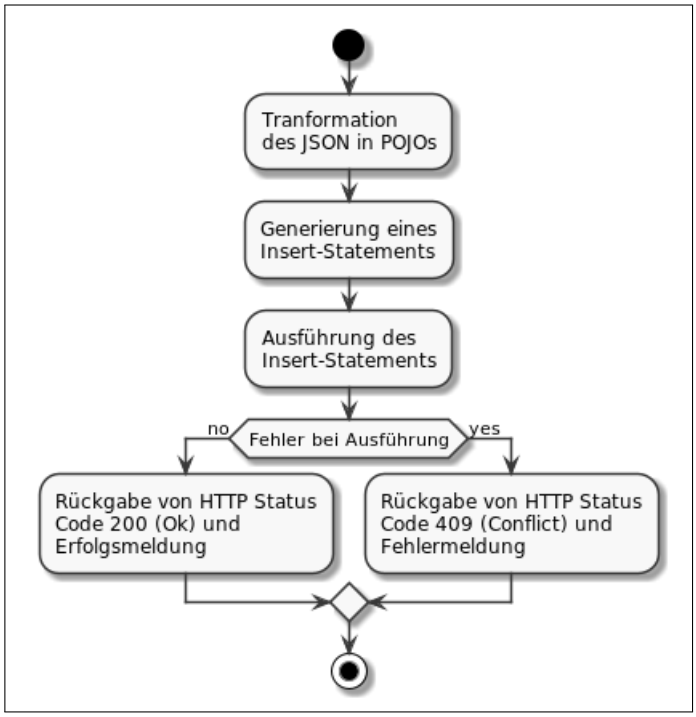

Abbildung 10: Aktivitätsdiagramm *Integration eines einzelnen Datensatzes* (Quelle: Eigene Darstellung)

Zeile 3 bis 10 erstreckt sich das Array *values*, welches wiederum Arrays mit JSON-Objekten enthält. Jedes Array beinhaltet einen Datensatz. Die einzelnen Objekte (Zeile 6, 7 und 8) innerhalb des Arrays repräsentieren die Spalte einer Zeile. Jedes Objekt enthält somit einen Spaltennamen (Schlüssel *columnName*) und den jeweiligen Wert mit einem Schlüssel *value*. Eine Spalte kann einen Wert vom Typ Integer, Float, Double oder Varchar enthalten.
Die beiden im Folgenden beschriebenen Aktivitäten *Integration von multiplen Datensätzen* sowie *Zeilenweise Integration von multiplen Datensätzen* geben die Funktionsweise der Endpoints /bulkUpload und /bulkUploadByRow wieder. Abb. 11 auf S. 64 enthält die Aktivität *Integration von multiplen Datensätzen*. Die Aktivität startet mit der Transformation des JSON-Objekts in POJOs. Im Anschluss daran wird ein Bulk-Insert-Statement generiert, wel-

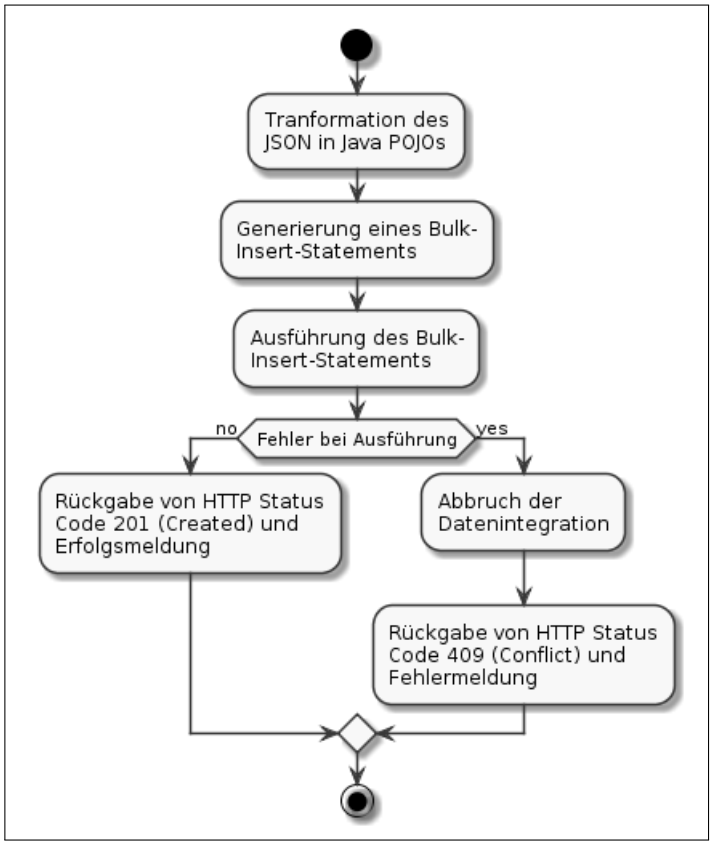

Abbildung 11: Aktivitätsdiagramm *Integration von multiplen Datensätzen* (Quelle: Eigene Darstellung)

```
1 {
2   [
3     {"Meldung 1"},
4     {"Meldung 2"},
5     ...
6     {"Meldung X"}
7   ]
8 }
```

Listing 4: JSON-Format von Erfolgs- und Fehlermeldungen (Quelle: Eigene Darstellung)

ches im nächsten Schritt ausgeführt wird. Bei einer fehlerfreien Ausführung werden der Status-Code 201 sowie eine Erfolgsmeldung zurückgegeben. Im Falle eines auftretenden Fehlers wird die gesamte Datenintegration abgebrochen und der Status-Code 409 sowie eine Fehlermeldung werden zurückgegeben. Die Fehlermeldung wird wiederum durch ein JSON-Objekt repräsentiert, welches Listing 4 auf S. 65 zu entnehmen ist. Zeile 1 und 8 enthalten die öffnende und schließende Klammer des JSON-Objekts. Von Zeile 2 bis 7 erstreckt sich das Array mit den einzelnen Objekten (Zeile 3 bis 6). Diese enthalten die einzelnen Meldungen. Das Array kann beliebig viele Meldungen enthalten. Abhängig vom Auftreten eines Fehlers, wird durch das JSON-Objekt eine Fehler- oder eine Erfolgsmeldung zurückgegeben. Das Array des Objekts enthält dabei nur ein Element, da nur eine Meldung zurückgegeben wird.

Die Aktivität *Zeilenweise Integration von multiplen Datensätzen* ist Abb. 12 auf S. 66 zu entnehmen. Der erste Schritt der Aktivität besteht aus der Transformation des JSON-Objekts in POJOs. Nach der erfolgreichen Transformation werden für jeden Datensatz, welcher durch ein POJO repräsentiert wird, ein Insert-Statement für einen einzelnen Datensatz gebildet. Anschließend wird das Statement ausgeführt. Bei einer erfolgreichen Ausführung des Statements wird eine Erfolgsmeldung in einer Liste hinterlegt. Entsprechend wird im Fehlerfall eine Fehlermeldung in dieser Liste hinterlegt. Dieses Vorgehen wird für alle POJOs und die durch sie repräsentierten Datensätze wiederholt. Sind keine weiteren POJOs vorhanden, werden der Status-Code 200 sowie eine Liste mit Erfolgs- und Fehlermeldungen zurückgegeben. Das

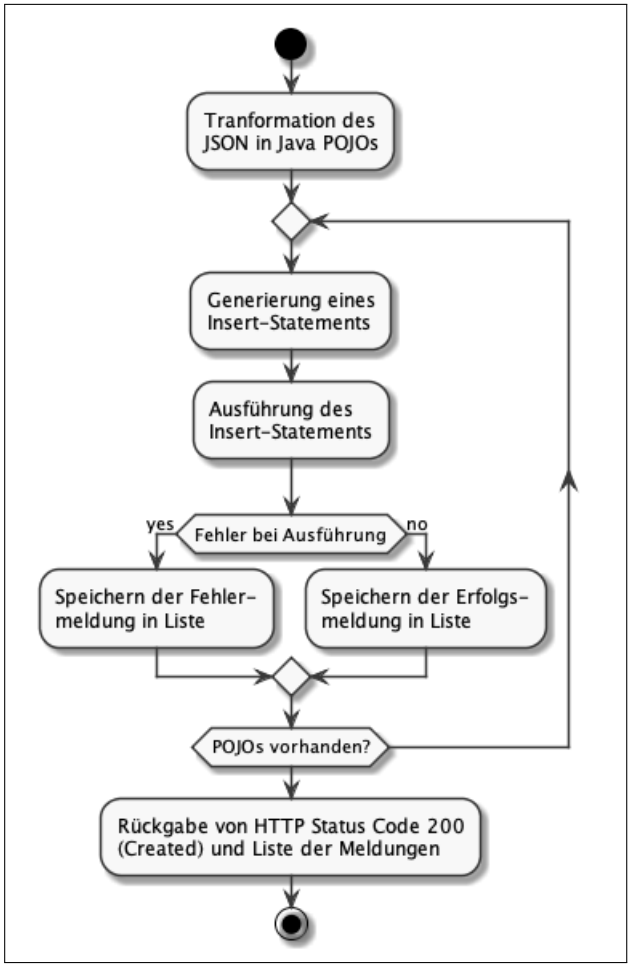

Abbildung 12: Aktivitätsdiagramm *Zeilenweise Integration von multiplen Datensätzen* (Quelle: Eigene Darstellung)

Format des zurückgegebenen JSON-Objekts entspricht dem aus Listing 4 auf S. 65.
Eine Übersicht über die Implementierung der Funktionalitäten für die Integration eines einzelnen Datensatzes sowie multipler Datensätze ist dem Klassendiagramm in Abb. 13 auf S. 68 zu entnehmen. Dieses enthält die Klassen *UploadController*, *UploadJsonMapper*, *BulkUploadJsonMapper*, *DataInsertException*, *UploadService* und *InsertDao* aus dem Package *org.dataint.server.insert* sowie die Klassen *ColumnValue* aus dem Package *org.dataint.server.model* und *SqlStatementCreator* aus dem Package *org.dataint.server.utils*. Den Ausgangspunkt für die Funktionalitäten der Datenintegration bildet eine Instanz der Controller-Klasse UploadController. In dieser Klasse sind die drei Endpoints /uploadData, /bulkUpload und /bulkUploadByRow definiert und mit den entsprechenden Methoden verknüpft. Je nach aufgerufenem Endpoint findet die Transformation des Json-Objekts in Instanzen der Klasse UploadJsonMapper oder BulkUploadJsonMapper statt. So wird ein einzelner Datensatz, der über den Endpoint /uploadData empfangen wird, in eine Instanz der Klasse UploadJsonMapper transformiert. Die Transformation von multiplen Datensätzen über die beiden Endpoints /bulkUpload und /bulkUploadByRow findet in eine Instanz der Klasse BulkUploadJsonMapper statt. Eine Instanz der Klasse UploadJsonMapper enthält neben dem Tabellennamen keine, eine oder beliebig viele Instanzen der Klasse ColumnValue, die eine Abbildung eines Tabellenspalteneintrags darstellt. Eine Instanz der Klasse BulkUploadJsonMapper hingegen enthält neben dem Tabellennamen eine Liste, in der wiederum Listen mit beliebig vielen Instanzen der Klasse ColumnValue enthalten sein können. Die Transformation wird in beiden Fällen durch das verwendete Framework[52] vorgenommen. Nach erfolgter Transformation delegiert der UploadController die Persistierung der Daten an eine Instanz der Klasse UploadService, welche ihrerseits an eine Instanz der Klasse InsertDao delegiert. Die Service-Klasse wird hier als Zwischenschicht verwendet, um DAO-Instanzen für beliebige datenhaltende Systeme einfügen zu können. Über die InsertDao-Klasse findet die eigentliche Persistierung der Daten in der Datenbank statt. Die notwendi-

[52]Siehe Kapitel 3.3 ab S. 49.

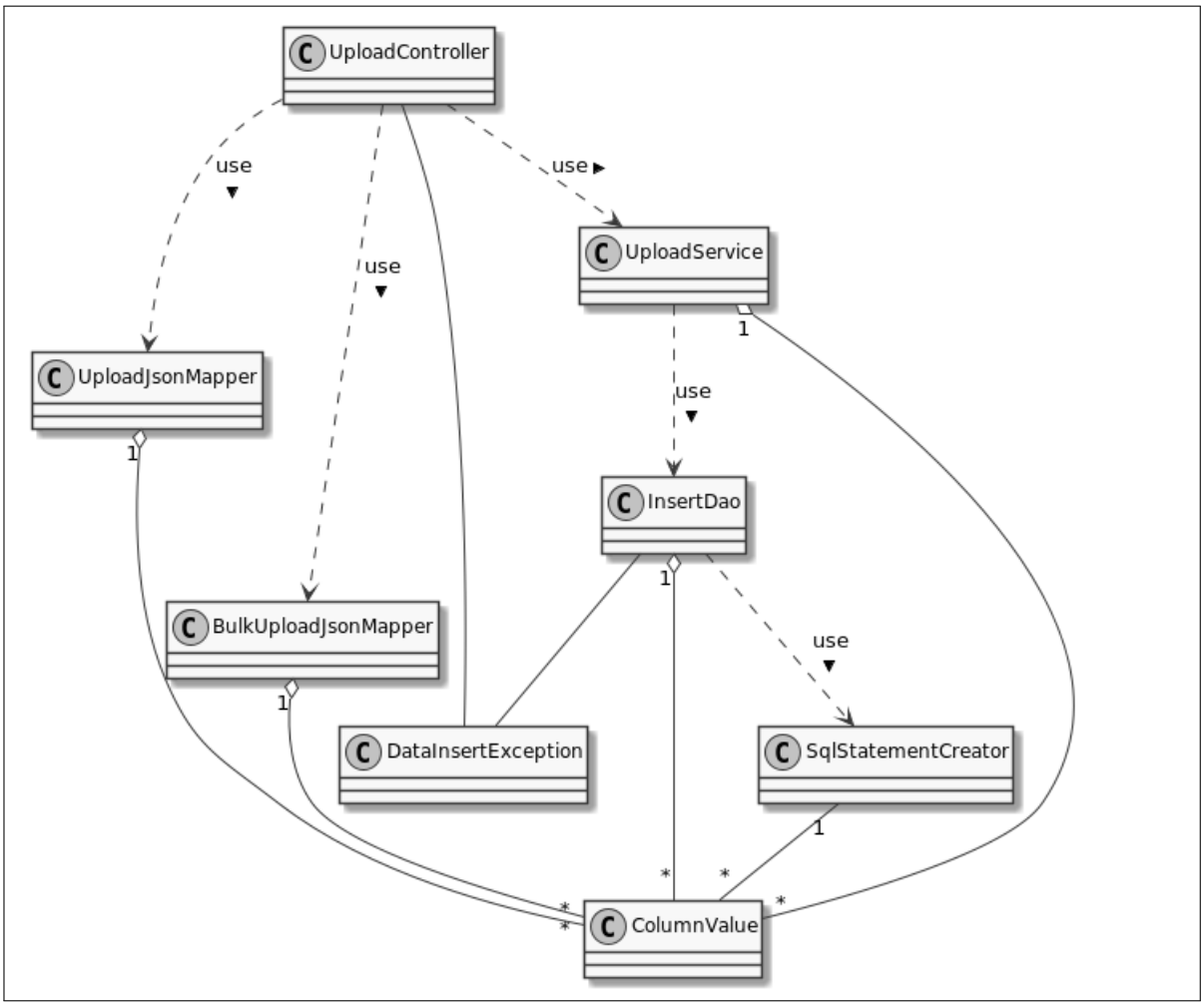

Abbildung 13: Klassendiagramm *Daten-Integration* (Quelle: Eigene Darstellung)

gen Insert-Statements werden der InsertDao-Instanz durch eine Instanz der Klasse SqlStatementCreator zur Verfügung gestellt. Diese Klasse ist in der Lage, sowohl ein Insert-Statement für einen einzelnen Datensatz als auch ein Insert-Statement für multiple Datensätze zu generieren. Für die zeilenweise Integration von multiplen Datensätzen geht eine InsertDao-Instanz wie folgt vor: Es findet eine Iteration über eine Liste mit allen Datensätzen statt. Für jeden Datensatz wird mit Hilfe der SqlStatementCreator-Instanz ein Insert-Statement für diesen einzelnen Datensatz gebildet. Die InsertDao-Instanz führt dieses Statement im Anschluss aus. Durch dieses Vorgehen bei der zeilenweisen Integration von multiplen Datensätzen mit einzelnen Insert-Statements pro Datensatz wird gewährleistet, dass bei einem Fehlerauftritt sowohl ein Logging[53] für einen einzelnen Datensatz möglich ist und zum anderen der Datenintegrationsvorgang nicht abgebrochen[54] werden muss, falls bei der Integration eines einzelnen Datensatzes ein Fehler auftritt.

4.3.2 CSV-Client

Der CSV-Client[55] entspricht dem Integrationswerkzeug[56] des Datenbeschaffungsbereichs und ermöglicht die Integration von einem oder mehreren Datensätzen aus beliebigen CSV-Dateien. Der Client erhält alle für eine Datenintegration notwendigen Informationen aus einer Extensible Markup Language (XML)-Konfigurationsdatei, deren Format Listing 5 auf S. 70 entnommen werden kann. Zeile 1 und 17 enthalten das umschließende *config*-Tag, innerhalb dessen alle Konfigurationsinformationen über den Client enthalten sind. Die Informationen für die URL des zu verwendenden Endpoints[57] sind dem öffnenden und schließenden Tag *url* in Zeile 2 beziehungsweise 7 des Listings zu entnehmen. Eine URL setzt sich somit zusammen aus dem Protokoll $*protocol* (Tag *protocol*, Zeile 3) Hostnamen $*hostname* (Tag *hostname*, Zei-

[53] Umsetzung der Anforderung *Logging* aus Kapitel 4.1 ab S. 52

[54] Umsetzung der Anforderung *Integration von multiplen Datensätzen* aus Kapitel 4.1 ab S. 52

[55] Umsetzung der Anforderung *Integration von CSV-Dateien* aus Kapitel 4.1 ab S. 52

[56] Siehe Kapitel 3.1 auf S. 41.

[57] Siehe Kapitel 4.3.1 ab S. 61 u. S. 62.

```
1  <config>
2    <url>
3      <protocol>${protocol}</protocol>
4      <hostname>${hostname}</hostname>
5      <port>${port}</port>
6      <path>${path}</path>
7    </url>
8    <csv-file-path>pfad_zur_datei.csv</csv-file-path>
9    <tablename>tabellenname</tablename>
10   <columnmappings>
11     <mapping>
12       <csvcolumnname>CSV-Spaltenname</csvcolumnname>
13       <tablecolumnname>Db-Spaltenname</tablecolumnname>
14     </mapping>
15     ...
16   </columnmappings>
17 </config>
```

Listing 5: Konfiguration des CSV-Uploads (Quelle: Eigene Darstellung)

le 4), dem Port $*port* (Tag *port*, Zeile 5) sowie dem Pfad $*path* (Tag *path*, Zeile 6). Aus Zeile 8 (Tag *csv-file-path*) kann der vollständige Dateipfad der einzulesenden CSV-Datei entnommen werden. Zeile 9 weist das Tag *tablename* auf. Der hier eingetragene Wert bezeichnet den Namen der Tabelle, in welche die Datenintegration erfolgt. Zeile 10 und 16 enthalten das öffnende beziehungsweise schließende Tag *columnmappings*. Innerhalb der beiden Tags können beliebig viele Mapping-Einträge vorgenommen werden. Ein solcher Mapping-Eintrag ermöglicht das Mapping einer Spalte in der zu integrierenden CSV-Datei auf eine bestimmte Spalte der Datenbank, in welche die Datensätze zu integrieren sind. Mapping-Einträge sind vorzunehmen, wenn sich Quellspalten einer CSV-Datei von den jeweiligen Zielspalten der Datenbanktabelle namentlich unterscheiden. Ein entsprechender Eintrag (Tag *mapping*) kann Zeile 11 bis 14 entnommen werden. Innerhalb eines Mapping-Eintrags wird über das Tag *csvcolumnname* die Spalte innerhalb der CSV-Datei (Zeile 12), über das Tag *tablecolumnname* der Datenbankspaltenname (Zeile 13) angegeben.

Die Konfigurationsdatei bildet die Grundlage für den Programmablauf des CSV-Clients. Der Programmablauf kann der Aktivität in Abb. 14 auf S. 71 entnommen werden. Der Aktivität nach wird zunächst die Konfigurationsdatei eingelesen. Falls keine solche existiert, wird die Aktivität beendet. Bei

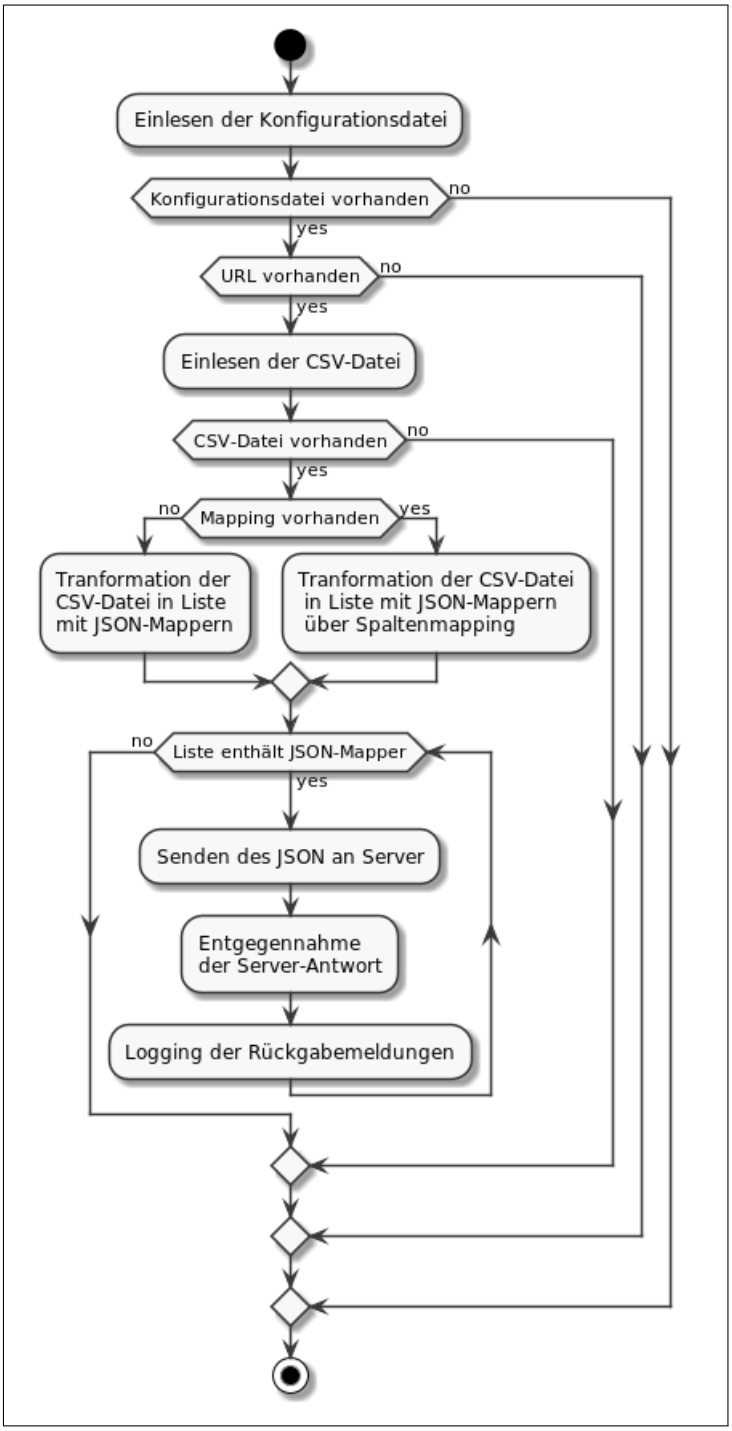

Abbildung 14: Aktivitätsdiagramm *CSV-Upload* (Quelle: Eigene Darstellung)

Existenz einer Konfigurationsdatei wird überprüft, ob aus dieser eine URL für einen Endpoint entnommen werden kann. Falls nicht wird die Aktivität beendet. Andernfalls wird die in der Konfigurationsdatei angegebene CSV-Datei eingelesen. Misslingt dies, so wird die Aktivität beendet. Im Erfolgsfall wird überprüft, ob Mapping-Einträge vorhanden sind. Sind solche vorhanden, wird unter Berücksichtigung des Mappings eine Transformation des CSV-Formats in JSON-Mapper vollzogen. Sind der Konfigurationsdatei keine Mapping-Einträge zu entnehmen, erfolgt diese Transformation der Datensätze aus der CSV-Datei ohne ein Mapping von Spalten, ansonsten mit einem Spaltenmapping. Im Anschluss daran werden die einzelnen JSON-Mapper durch das verwendete Framework[58] in JSON-Objekte von multiplen Datensätzen[59] umgewandelt, an den Server gesendet und die Antwort des Servers entgegengenommen. Anschließend werden die Rückgabemeldungen[60] des Servers geloggt. Diese Vorgänge werden so lange wiederholt, wie JSON-Mapper vorhanden sind. Die Aufteilung der CSV-Datei in viele JSON-Mapper erfolgt aus der Notwendigkeit, den Body eines einzelnen Requests größenmäßig zu beschränken. Ist ein Request nämlich zu groß, würde die Server-Instanz einen Request ablehnen. Sind keine weiteren JSON-Mapper vorhanden, endet die Aktivität.

Abb. 15 auf S. 73 zeigt das Klassendiagramm des CSV-Clients. Die Klassen befinden sich im Package org.dataint.client. Den initialen Startpunkt des CSV-Clients bildet eine Instanz der Klasse *App*. Den ersten Schritt, das Einlesen der Konfigurationsdatei[61], delegiert die App-Instanz an eine Instanz der Klasse *ConfigFileExtractor*. Die ConfigFileExtractor-Instanz speichert alle eingelesenen Informationen in einer Instanz der Klasse *AppConfig*. Auf diese Instanz greift auch die App-Instanz zu. Das Ein- und Auslesen der CSV-Datei mit den eigentlichen Datensätzen delegiert die App-Instanz wiederum an eine *CSVReader*-Instanz. Diese Reader-Instanz benötigt für das Ein- und Auslesen wiederum die AppConfig-Instanz mit den gespeicherten Konfigurations-

[58] Siehe Kapitel 3.3 ab S. 49.
[59] Siehe Listing 3 in Kapitel 4.3.1 auf S. 62.
[60] Siehe Listing 4 in Kapitel 4.3.1 auf S. 65.
[61] Siehe Listing 5 auf S. 70.

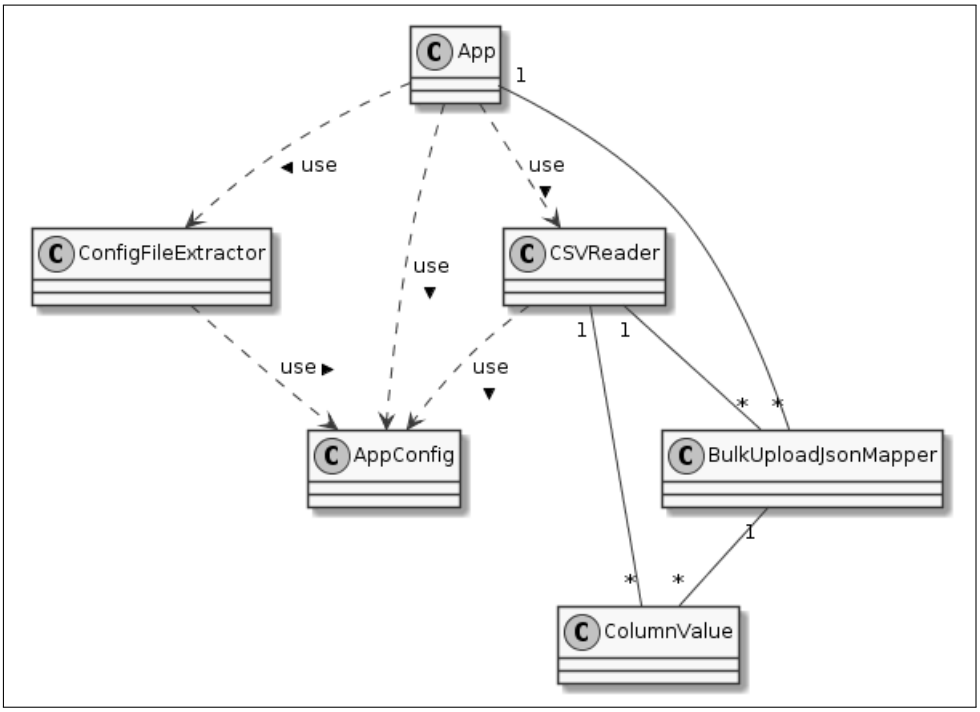

Abbildung 15: Klassendiagramm *CSV-Client* (Quelle: Eigene Darstellung)

informationen. Während des Einlesevorgangs transformiert die CSVReader-Instanz die eingelesenen Datensätze aus der CSV-Datei in Instanzen der Klasse *BulkUploadJsonMapper* unter Berücksichtigung eventueller Mapping-Einträge. Eine solche Instanz besitzt keine, eine oder viele *ColumnValue*-Instanzen. Eine ColumnValue-Instanz repräsentiert dabei eine Datenbankspalte innerhalb der Applikation und weist den Spaltennamen sowie -wert auf. Nach erfolgreicher Transformation gibt die CSVReader-Instanz die eingelesenen Daten als BulkUploadJsonMapper-Instanzen an die App-Instanz zurück. Mit Hilfe des verwendeten Frameworks[62] werden die BulkUploadJsonMapper-Instanzen in JSON-Strings umgewandelt und diese an den Server gesendet. Die Meldungen aus den verschiedenen Responses des Servers werden abschließend in der App-Instanz geloggt.

[62]Siehe Kapitel 3.3 ab S. 49.

4.4 Anwendung

Das im Rahmen der vorliegenden Arbeit entwickelte System wurde innerhalb dieser Arbeit in verschiedenen Schritten verwendet. Diese Schritte waren:

1. Konfiguration der Data-Server-Instanzen
2. Anlegen der Tabellen in der Basisdatenbank
3. Integration der Datensätze in die Basisdatenbank
4. Aggregation der Datensätze innerhalb der Basisdatenbank
5. Export der aggregierten Datensätze als CSV-Datei
6. Anlegen der Tabellen in der Ableitungsdatenbank
7. Upload der aggregierten und modifizierten Datensätze in die Ableitungsdatenbank

Im weiteren Verlauf dieses Unterkapitels werden die einzelnen Schritte beschrieben.

Als erster Schritt wurden die beiden Data-Server-Instanzen Integration und Delivery aufgesetzt. Die beiden Instanzen werden über die gleiche jar-Datei, jedoch mit unterschiedlichen Profilen aufgerufen. Die Konfigurationen der beiden Profile sind in den Property-Dateien *application-integration.properties* (Listing 6, S. 74) sowie *application-delivery.properties* (Listing 7, S. 75) hinterlegt und unterscheiden sich unter anderem sowohl durch den Datenbanknamen (Zeile 2) als auch durch den Port (Zeile 4).

```
1 ...
2 spring.datasource.url=jdbc:mysql://localhost:3306/
    integration?serverTimezone=UTC
3 ...
4 server.port=2048
```

Listing 6: Port- und Datenbankkonfiguration der Data-Server-Instanz Integration (Quelle: Eigene Darstellung)

```
1 ...
2 spring.datasource.url=jdbc:mysql://localhost:3306/
     delivery?serverTimezone=UTC
3 ...
4 server.port=2050
```

Listing 7: Port- und Datenbankkonfiguration der Data-Server-Instanz Delivery (Quelle: Eigene Darstellung)

Der Datenbankname für die Integrations-Instanz des Data-Servers lautet *integration*, für die Delivery-Instanz entsprechend *delivery*.

Der zweite Schritt bestand im Anlegen der Tabellen, die für die Integration der Datensätze vorhanden sein müssen. Der Vorgang kann Abb. 16 auf S. 75 entnommen werden. Über einen *Client*[63] wurden die Tabellenkonfigura-

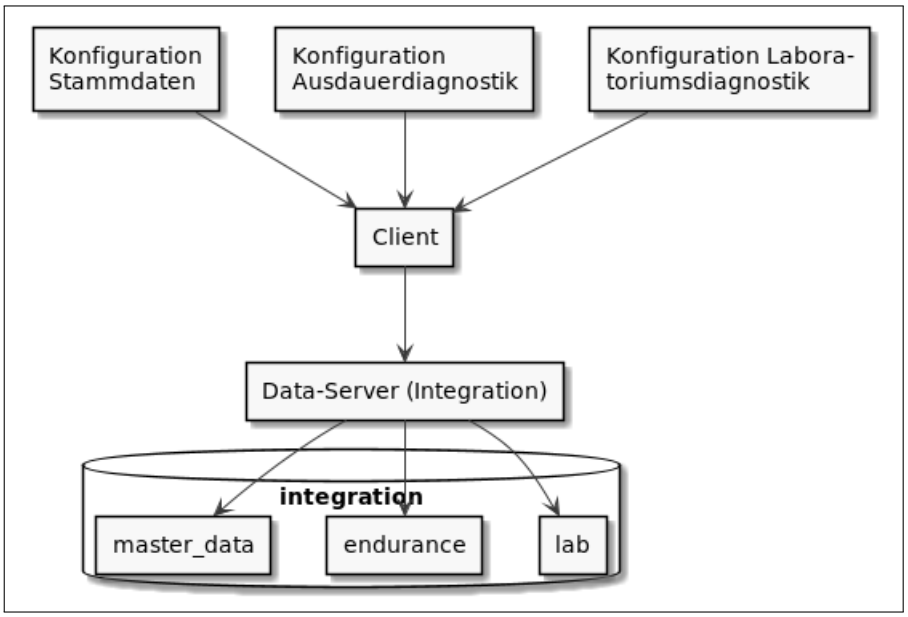

Abbildung 16: Systemanwendung für das Erstellen der Tabellen innerhalb der Basisdatenbank (Quelle: Eigene Darstellung)

tionen *Konfiguration Stammdaten*[64], *Konfiguration Ausdauerdiagnostik*[65] sowie *Konfiguration Laboratoriumsdiagnostik*[66] über einzelne POST-Requests

[63] Siehe Kapitel 3.3 ab S. 49.
[64] Siehe Listing 8 in Kapitel A.1.1 ab S. 154.
[65] Siehe Listing 9 in Kapitel A.1.1 ab S. 155.
[66] Siehe Listing 10 in Kapitel A.1.1 ab S. 156.

als JSON-Objekte an die Instanz *Data-Server (Integration)* gesendet. Die Data-Server-Instanz legte für jeden erfolgreich ausgeführten Request die entsprechenden Tabellen *master_ data*, *endurance* beziehungsweise *lab* innerhalb der Basisdatenbank *integration* an. Die Data-Server-Instanz Integration sowie die Datenbank integration stellen den Integrationsbereich dar.[67]

Der dritte Schritt, die eigentliche Integration der Datensätze aus Stammdaten, Ausdauerdiagnostik und Laboratoriumsdiagnostik erfolgte im Anschluss an das Erstellen der Tabellen für die Data-Server-Instanz Integration und ist in Abb. 17 auf S. 76 dargestellt. Bei der Integration kam der *CSV-Data-*

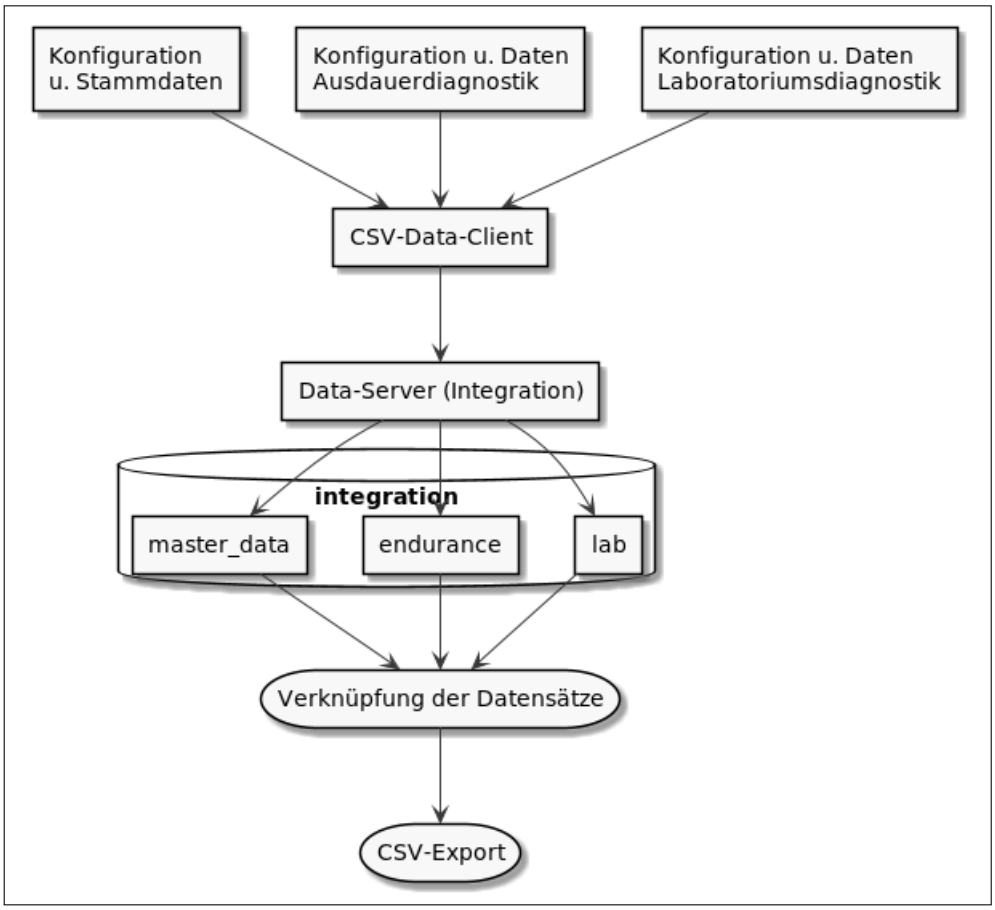

Abbildung 17: Systemanwendung für die Datenintegration und -aggregation (Quelle: Eigene Darstellung)

[67]Siehe Kapitel 3.1 ab S. 41.

Client zum Einsatz. Mit dessen Hilfe wurden jeweils die Konfigurationsdatei wie auch die CSV-Datei mit den Datensätzen geladen und über den Endpoint /bulkUploadByRow[68] der Data-Server-Instanz Integration an diese übergeben. Dieses Procedere wurde sowohl für die Stammdatensätze (*Konfiguration*[69] *u. Stammdaten*), die Datensätze der Ausdauerdiagnostik (*Konfiguration*[70] *u. Daten Ausdauerdiagnostik*) als auch für die Datensätze der Laboratoriumsdiagnostik (*Konfiguration*[71] *u. Daten der Laboratoriumsdiagnostik*) durchgeführt. Die Server-Instanz persistierte die jeweiligen Datensätze in den entsprechenden Tabellen *master_ data*, *endurance* und *lab* der Datenbank *integration*.

Nachdem die Datensätze von Stammdaten, Ausdauer- sowie Laboratoriumsdiagnostik in der Datenbank persistiert wurden, fand im vierten Schritt die Verknüpfung[72] dieser Datensätze statt. Datensätze mit Missings flossen nicht in die neuen Datensätze mit ein.[73]

Als fünfter Schritt wurden die entstandenen Datensätze wiederum als CSV-Datei aus der Datenbank integration exportiert und standen anschließend zur weiteren Bearbeitung innerhalb der Auswertung[74] zur Verfügung.

Vor und während der Auswertung wurden aggregierte und modifizierte Datensätze in die Ableitungsdatenbank geladen. Als Voraussetzung dafür wurden mit dem sechsten Schritt die Tabellen für diese Datensätze angelegt, wie in Abb. 18 auf S. 78 zu sehen ist. Dafür wurden zunächst über zwei GET-Requests die beiden Tabellenkonfigurationen *Konfiguration aggregierte Daten*[75] und *Konfiguration modifizierte Daten*[76] an die Instanz *Data-Server*

[68] Siehe Kapitel 4.3.1 ab S. 62.
[69] Siehe Listing 13 in Kapitel A.2 ab S. 160.
[70] Siehe Listing 14 in Kapitel A.2 ab S. 160.
[71] Siehe Listing 15 in Kapitel A.2 ab S. 161.
[72] Siehe Kapitel 3.1 ab S. 44.
[73] Umsetzung der Anforderung *Aggregation von Datensätzen* aus Kapitel 4.1 ab S. 53
[74] Siehe Kapitel 3.1 ab S. 43.
[75] Siehe Kapitel A.1 ab S. 157.
[76] Siehe Kapitel A.1 ab S. 158.

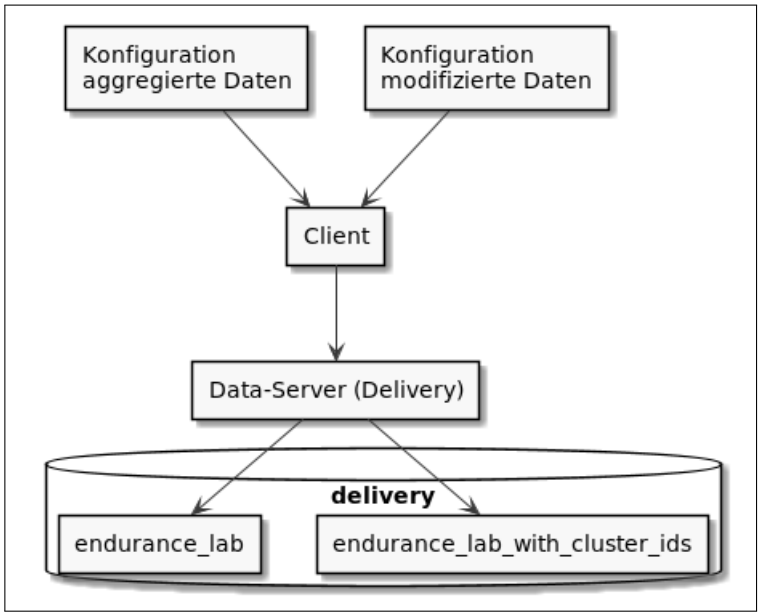

Abbildung 18: Systemanwendung für das Erstellen der Tabellen innerhalb der Ableitungsdatenbank (Quelle: Eigene Darstellung)

(Delivery) gesendet. Diese Instanz legt nach erfolgreichem Senden die beiden Tabellen *endurance_lab* und *endurance_lab_with_cluster_ids* innerhalb der Datenbank *delivery* an. Data-Server-Instanz und Datenbank entsprechen hier dem Auswertebereich des Systems.[77]

Der Upload von Datensätzen in die Ableitungsdatenbank entspricht dem siebten und damit letzten Schritt und kann Abb. 19 auf S. 79 entnommen werden. Auch beim Upload von Datensätzen in die Datenbank *delivery* kam der *CSV-Data-Client* zum Einsatz. Dieser entspricht dem Integrationswerkzeug[78] im Auswertebereich. Der Client las sowohl die Konfigurations- als auch die jeweilige CSV-Datei mit den Datensätzen ein. In der Abbildung sind diese als *Konfiguration*[79] *und modifizierte Datensätze* und *Konfiguration*[80] *und aggregierte Datensätze* gekennzeichnet. Der CSV-Data-Client sendete die jeweiligen Datensätze über den Endpoint /bulkUploadByRow an die Instanz *Data-*

[77]Siehe auch Kapitel 3.1 ab S. 43.
[78]Siehe auch Kapitel 3.1 ab S. 41.
[79]Siehe Listing 16 in Kapitel A.2.2 ab S. 161.
[80]Siehe Listing 17 in Kapitel A.2.2 ab S. 162.

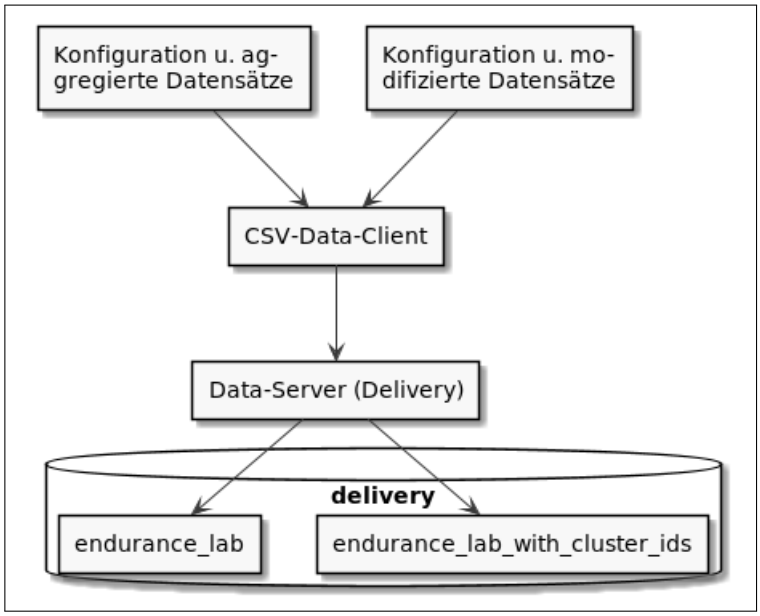

Abbildung 19: Systemanwendung für den Upload (Quelle: Eigene Darstellung)

Server (Delivery). Diese persistierte die Datensätze in den entsprechenden Tabellen *endurance_lab* beziehungsweise *endurance_lab_with_cluster_id*. Bei den Datensätzen in der Tabelle endurance_lab handelt es sich um die aggregierten Datensätze, welche die berechneten Werte für den Parameter tlim enthalten. Die Datensätze in der Tabelle endurance_lab_with_cluster_ids beinhalten darüber hinaus die Clusterbezeichnung aus dem Clustering[81]. Im Anschluss standen die Datensätze für weitere Betrachtungen zur Verfügung.

[81]Siehe Kapitel 3.1 ab S. 44.

5 Exemplarische Ergebnisse: Beschreibung und Diskussion

Innerhalb des vorliegenden Kapitels werden die Ergebnisse des zweistufigen Datenmodells[82] bestehend aus einem hierarchischen Clustering[83] und einem Decision Tree[84] jeweils beschrieben und im sportwissenschaftlichen Kontext betrachtet und diskutiert.

5.1 Cluster und Gruppen

Abb. 20 auf S. 81 enthält die grafische Darstellung des aus dem Clustering resultierenden Dendogramms[85]. Auf der x-Achse sind Integer-Zahlen von 1 bis 58 als Repräsentation der einzelnen Individuen verzeichnet. Die y-Achse weist die Distanz auf, deren Skalierung fünfstellig ist und von 0 bis 40 reicht. Die cut-off Linie liegt bei einer Distanz von 39 und schneidet zwei vertikale Linien an der größten Distanz zwischen zwei Clustern von 17.7, so dass bei der Analyse von zwei Clustern ausgegangen werden kann. Aufgrund der jeweiligen Mittelwerte innerhalb der Cluster werden im weiteren Verlauf dieser Arbeit Cluster 1 als Cluster low und Cluster 2 als Cluster high bezeichnet.

Tabelle 3 auf S. 81 gibt Aufschluss über die Mittelwerte von Cluster low. Die Tabelle enthält drei Spalten, je eine für den Parameter, die Einheit des Parameters sowie den arithmetischen Mittelwert (MW). Für die *tlim* beträgt der MW von Cluster low 19.3 min. Die *V4* ist mit 3.34 m/s beziffert. Die $rVO_2{}_{peak}$ weist einen Wert von 45.9 S/min auf. Der RQ_{peak} wird mit 1.07 angegeben. Die Hf_{max} weist einen Wert von 197 S/min auf. Die Lak_{peak} wird mit 5.74 mmol/l bestimmt. Die Individuen aus Cluster low weisen für den *Hb* im Mittel einen Wert von 13.3 g/dl auf.
Tabelle 4 auf S. 82 enthält die Mittelwerte von Cluster high und ist eben-

[82] Siehe Kapitel 3.1 ab S. 44.
[83] Siehe auch Kapitel 2.2.1 ab S. 22.
[84] Siehe auch Kapitel 2.2.2 ab S. 27.
[85] Siehe auch Kapitel 2.2.1 ab S. 26.

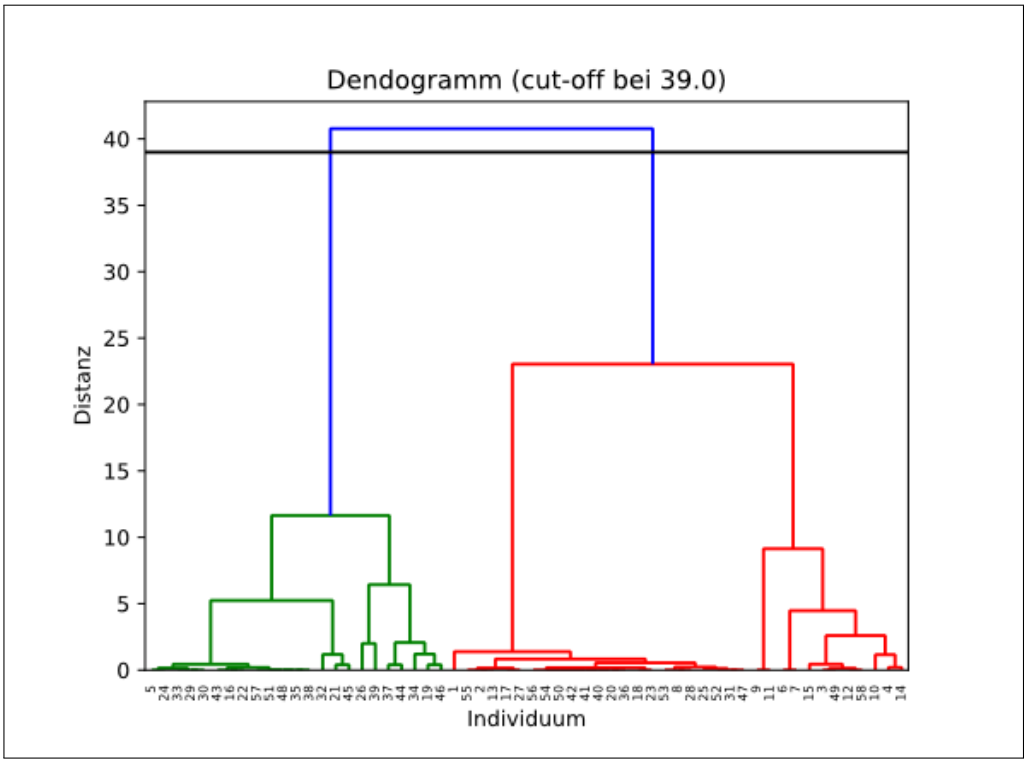

Abbildung 20: Graphische Darstellung des Dendogramms (Quelle: Eigene Darstellung)

Parameter	Einheit	MW
tlim	min	19.3
V4	m/s	3.34
rVO_{2peak}	ml/kg/min	45.9
RQ_{peak}	-	1.07
Hf_{max}	S/min	197
Lak_{peak}	mmol/l	5.74
Hb	g/dl	13.3

Tabelle 3: Parametermittelwerte von Cluster low (Quelle: Eigene Darstellung)

falls dreispaltig. Die Spalten enthalten wiederum die jeweiligen Parameter, deren Einheiten sowie die Mittelwerte, wie den Spaltenüberschriften zu entnehmen ist. Die Individuen aus Cluster high weisen im Mittel für die *tlim*

Parameter	Einheit	Mittelwert
tlim	min	27.0
V4	m/s	3.81
$rVO_{2_{peak}}$	ml/kg/min	53.0
RQ_{peak}	-	1.09
Hf_{max}	S/min	193
Lak_{peak}	mmol/l	7.15
Hb	g/dl	14.2

Tabelle 4: Parametermittelwerte von Cluster high (Quelle: Eigene Darstellung)

einen Wert von 27.0 min auf. 3.81 m/s beträgt der Wert für die *V4*. Für den Parameter $rVO_{2_{peak}}$ liegt ein Wert von 53.0 ml/kg/min vor. Auf 1.09 ist der Wert für den Parameter RQ_{peak} zu beziffern. Für die Hf_{max} kann der Tabelle ein Wert von 193 S/min entnommen werden. Der Parameter Lak_{peak} weist einen Wert von 7.15 mmol/l auf, 14.2 g/dl beträgt der Wert für den *Hb*.

Im Folgenden werden die Mittelwerte der beiden Cluster pro Parameter miteinander verglichen. Dazu zeigt Abb. 21 auf S. 83 ein Parallelkoordinatendiagramm, in dem die Mittelwerte der beiden Cluster für die einzelnen Parameter eingezeichnet sind. Die x-Achse des Diagramms enthält die untersuchten Parameter *tlim*, *V4*, $rVO_{2_{peak}}$, RQ_{peak}, Hf_{max}, Lak_{peak} und *Hb*. Auf den jeweiligen y-Achsen ist die entsprechende Skalierung der Parameter enthalten. Die Mittelwerte der beiden Cluster sind farblich unterschiedlich dargestellt. Die Werte von Cluster low sind schwarz, die von Cluster high grau markiert. Mit Hilfe des Diagramms werden im Folgenden die Wertedifferenzen bei den verschiedenen Parametern zwischen Cluster low und Cluster high veranschaulicht und beschrieben. Dem Diagramm ist ganz klar zu entnehmen, dass die Werteausprägungen von Cluster low für die einzelnen Parameter im Mittel niedriger sind als die von Cluster high mit Ausnahme des Parameters Hf_{max}.

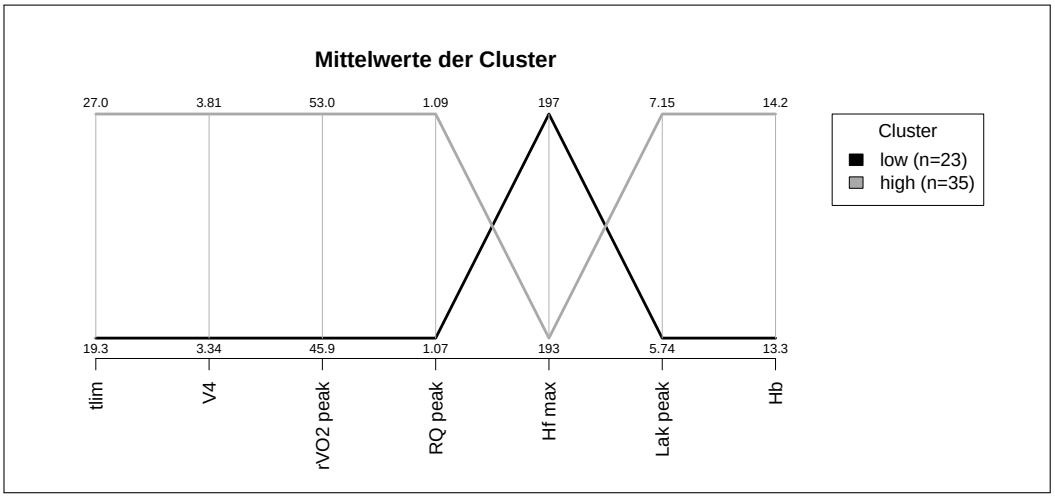

Abbildung 21: Mittelwerte der Cluster für die untersuchten Parameter (Quelle: Eigene Darstellung)

Diesen Umstand spiegelt die Namensgebung der beiden Cluster wider.
Im Folgenden werden die jeweiligen Differenzen für die einzelnen Parameter beschrieben. Für die tlim liegt eine Differenz von 6.7 min vor. Die V4 weist eine Differenz von 0.47 m/s auf. Für die $rVO_{2_{peak}}$ ist eine Differenz von 7.1 ml/kg/min zu verzeichnen. Die Differenz des RQ_{peak} beträgt 0.02 und ist somit zu vernachlässigen. Die Hf_{max} ist der einzige Parameter, bei dem die Individuen von Cluster high im Mittel einen niedrigeren Wert als die Individuen von Cluster low aufweisen. Die Differenz beträgt hier 4 S/min. Des Weiteren weist der Parameter Lak_{peak} eine Differenz von 1.41 mmol/l zwischen den Mittelwerten von Cluster low und Cluster high auf. Schließlich ist für den Parameter Hb eine Differenz von 0.9 g/dl zu erkennen.

Im weiteren Verlauf wird die Struktur der beiden gefundenen Cluster bezüglich Geschlecht, Alter und Sportart beschrieben.
Dazu zeigt Abb. 22 auf S. 84 zunächst die Geschlechterverteilung von Cluster low und Cluster high anhand eines Balkendiagramms. Auf der x-Achse des Diagramms sind die beiden Cluster eingetragen. Jedes Cluster wird durch jeweils einen schwarzen Balken für männliche und einen grauen Balken für weibliche Individuen repräsentiert. Die y-Achse enthält die Anzahl an In-

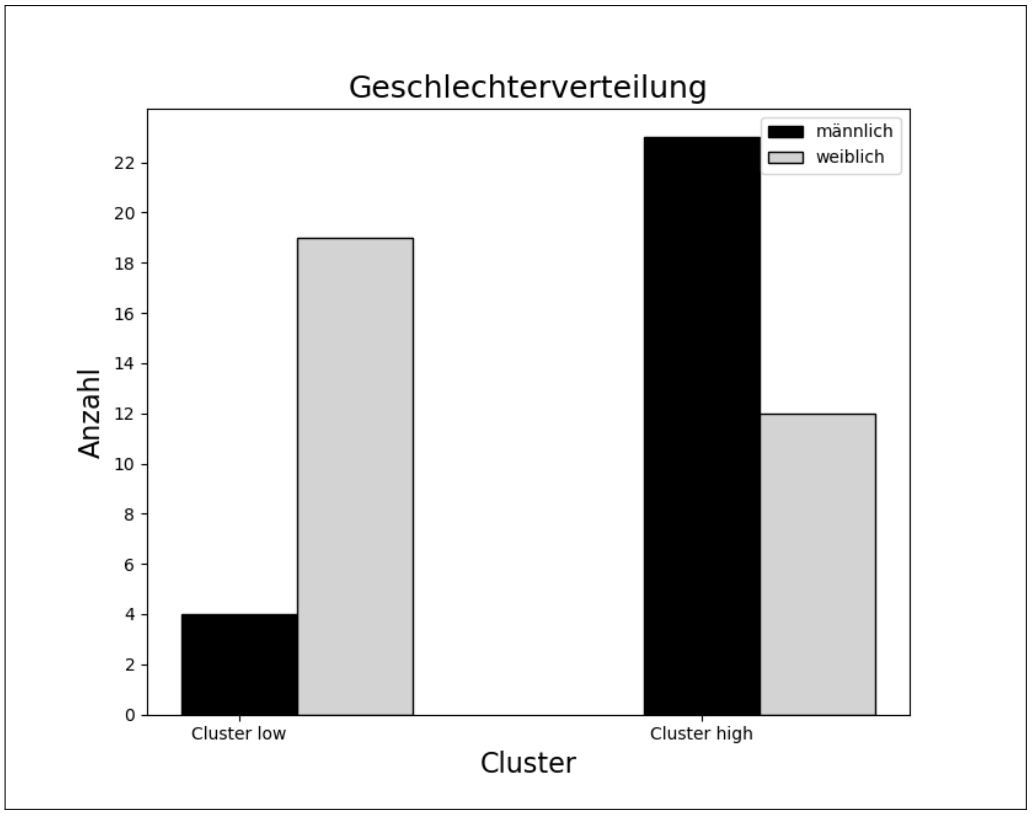

Abbildung 22: Balkendiagramm *Geschlechterverteilung* (Quelle: Eigene Darstellung)

dividuen, die durch eine zweistufige Skala mit Werten zwischen 0 und 22 angegeben sind. Cluster low kann somit in 4 männliche und 19 weibliche Individuen unterteilt werden. Cluster high hingegen sind 23 männliche und 12 weibliche Individuen zugeordnet.
Aus den Balkendiagrammen in Abb. 23 auf S. 85 und Abb. 24 auf S. 86 kann jeweils die Altersverteilung von Cluster low und Cluster high entnommen werden. Die Diagramme enthalten auf der x-Achse die Altersstufen der Individuen in Jahren. Die Altersstufen betragen 14, 15, 16, 17, 18, 20 und 25 Jahre. Auf der y-Achse ist die Anzahl der Individuen eingetragen. Die Anzahl männlicher Individuen ist schwarz, die der weiblichen Individuen ist grau gefärbt. Wie Abb. 23 zu entnehmen ist, sind Individuen im Alter von 14 bis 18 Jahren in Cluster low vertreten. Dem Cluster werden für das Alter von 14

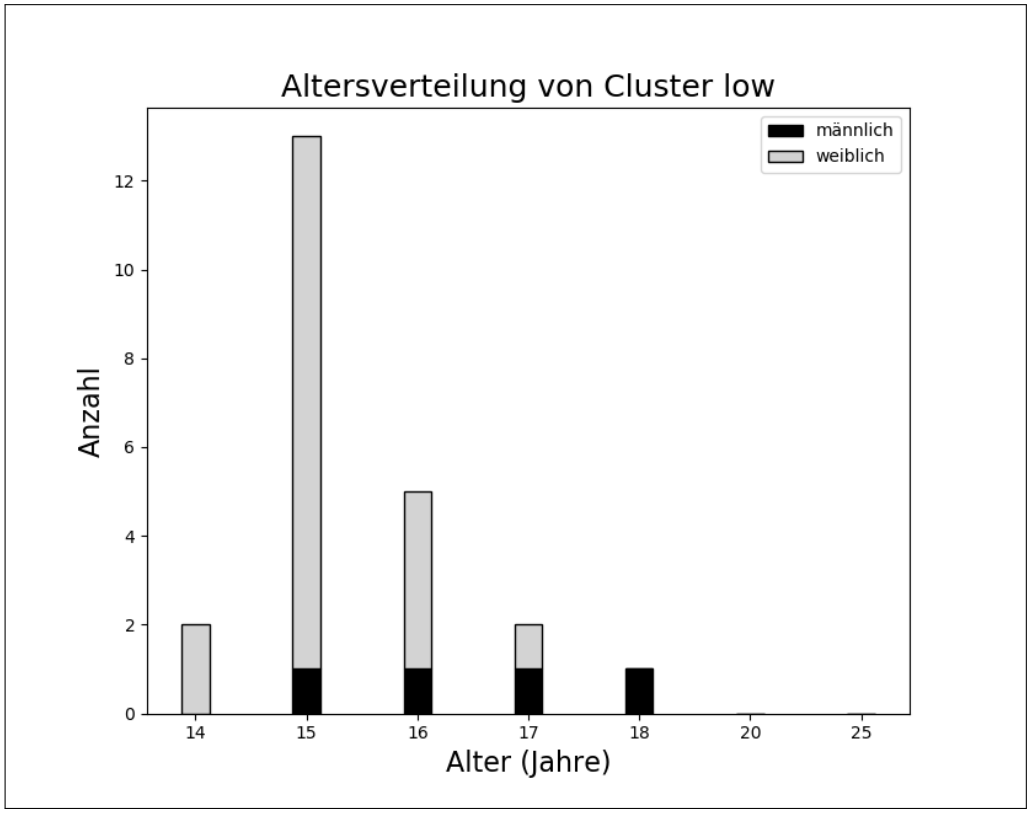

Abbildung 23: Balkendiagramm *Altersverteilung von Cluster low* (Quelle: Eigene Darstellung)

Jahren 2 weibliche Individuen zugeordnet. Des Weiteren enthält Cluster low

1 männliches und 12 weibliche Individuen mit 15 Jahren. In der Altersklasse von 16 Jahren sind 1 männliches und 4 weibliche Individuen, in der von 17 Jahren jeweils 1 männliches und 1 weibliches Individuum und in der Klasse von 18 Jahren 1 männliches Individuum vorhanden. Die Altersklassen von 20 und 25 Jahren sind in Cluster low nicht vertreten. Anhand von Abb. 24 kann die Altersverteilung für Cluster high ermittelt werden. Sie weist Individuen in den Altersstufen von 15 bis 25 Jahren auf. Für Cluster high liegen

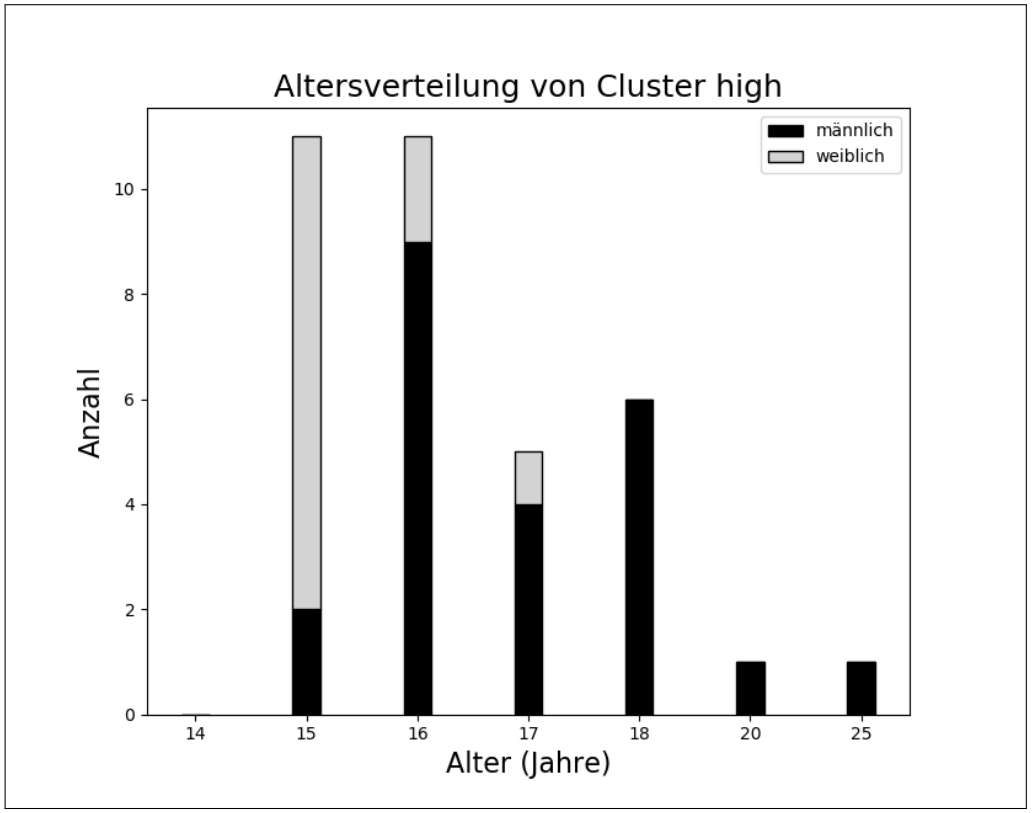

Abbildung 24: Balkendiagramm *Altersverteilung von Cluster high* (Quelle: Eigene Darstellung)

keine Individuen in der Altersklasse von 14 Jahren vor, in der Klasse von 15 Jahren 2 männliche und 9 weibliche. 9 männliche und 2 weibliche Individuen sind der Altersklasse von 16 Jahren zuzuordnen. Für die Altersklasse von 17 Jahren liegen vier männliche Individuen und ein weibliches Individuum vor. In den übrigen Alterklassen 18, 20 und 25 Jahre befinden sich nur männliche Individuen. In der Klasse von 18 sind dies 6 Individuen und für die Alters-

klassen 20 sowie 25 Jahre jeweils 1 Individuum.
Abb. 25 auf S. 87 zeigt die Sportartenverteilung der Cluster low und high. Auf der x-Achse sind die Sportarten der Cluster als Balken eingetragen. Die

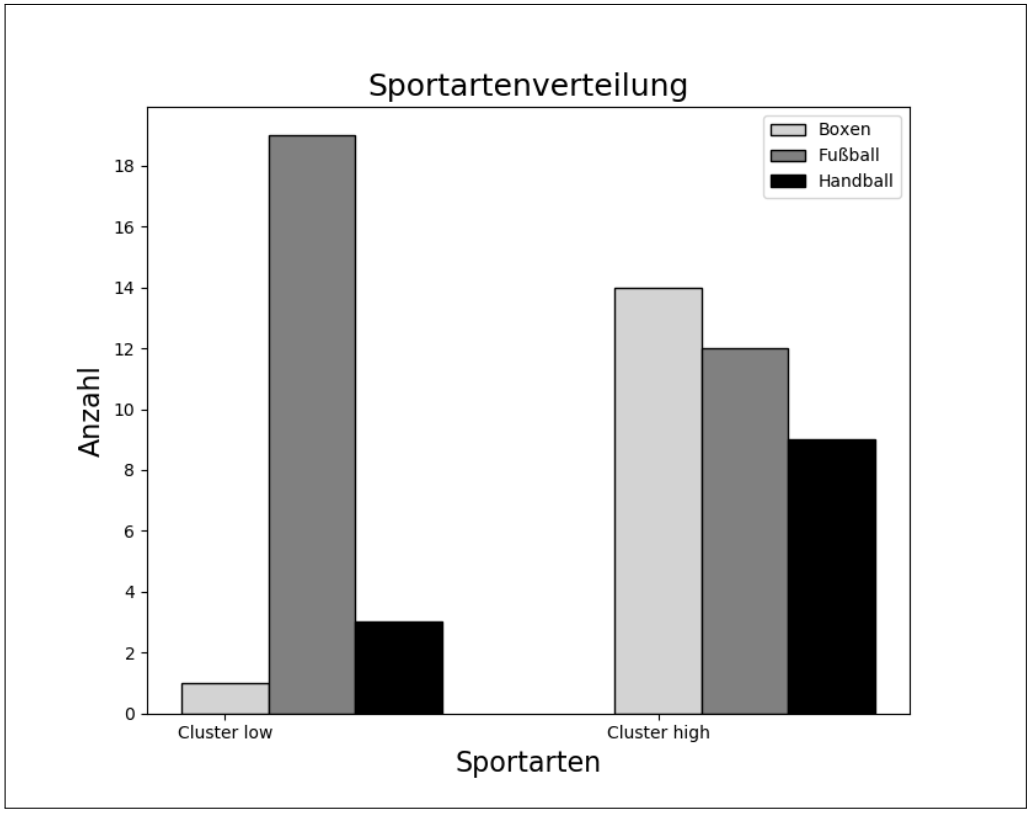

Abbildung 25: Balkendiagramm *Sportartenverteilung* (Quelle: Eigene Darstellung)

Sportarten selbst sind durch die Farben hellgrau für Boxen, dunkelgrau für Fußball und schwarz für Handball repräsentiert, wie der Legende entnommen werden kann. Auf der y-Achse ist die Anzahl an Indiviuduen mit einer zweistufigen Skala von 0 bis 18 zu sehen. In Cluster low ist zunächst ein Individuum der Sportart Boxen vertreten, neunzehn Individuen der Sportart Fußball und fünf Individuen der Sportart Handball. Cluster high weist vierzehn Individuen aus der Sportart Boxen, zwölf aus der Sportart Fußball und neun aus der Sportart Handball auf.

Für eine weitere, tiefergehende Betrachtung der Ergebnisse werden die In-

dividuen in Gruppen eingeteilt. Diese Einteilung wird zunächst aufgrund des durchgeführten Clusterings vollzogen, um eine Trennung auf der Basis der Leistungsattribute zu gewährleisten. Die weitere Unterteilung der Gruppen erfolgt durch die Zuordnung zu den Sportarten, um eine potentielle sportartenspezifische Adaption der betrachteten physiologischen Parameter transparent zu machen.
Es existieren somit die sechs Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball*. Die Gruppe high-Boxen umfasst 14, die Gruppe high-Fußball 12 und die Gruppe high-Handball 9 Individuen. Der Gruppe low-Boxen ist 1 Individuum und den Gruppen low-Fußball sowie low-Handball sind 19 beziehungsweise 3 Individuen zugeordnet.

Die Altersstruktur der Individuen kann den Boxplots aus Abb. 26 auf S. 89 sowie der Tabelle 5 auf S. 89 entnommen werden.
Die Abbildung enthält auf der x-Achse das Alter in Jahren im Bereich 14 bis 24 Jahre. Die Skala ist zweistufig. Auf der y-Achse sind die Gruppen high-Boxen, high-Fußball, high-Handball, low-Boxen, low-Fußball sowie low-Handball zu sehen. Der Median (Med) und der MW sind durch eine vertikale rote Linie beziehungsweise ein grünes Dreieck in den Boxplots kenntlich gemacht. Die Tabelle enthält in der ersten Zeile die Spaltennamen für Gruppenbezeichnung (Gruppe), Minimum (Min), Quartil 1 (Q1), Med, MW $\pm$ Standardabweichung (SD), Inter Quartil Range (IQR), Quartil 3 (Q3), Maximum (Max) sowie Ausreißer (Aus).
Der Wertebereich für die Gruppe high-Boxen liegt zwischen 16 und 25 (Aus) Jahren. Min und Q1 liegen bei 16 Jahren, der Med weist einen Wert von 17 Jahren auf. MW und SD werden mit 17.6 sowie $\pm$2.3 Jahren beziffert. Der IQR beträgt 2 Jahre. Q3 beträgt 18, das Max 20 Jahre.
Die Gruppe high-Fußball besitzt einen Wertebereich zwischen 15 und 17 Jahren. Min, Q1 und Med liegen bei 15 Jahren. MW und SD weisen Werte von 15.3 sowie $\pm$0.6 Jahren auf. Der IQR ist mit 0.2 Jahren verzeichnet. Das Q3 und das Max besitzen einen Wert von 15.2 Jahren. Für die Gruppe existieren zwei Aus von 16 und einer von 17 Jahren.
Der Wertebereich der Gruppe high-Handball liegt zwischen 15 und 18 Jah-

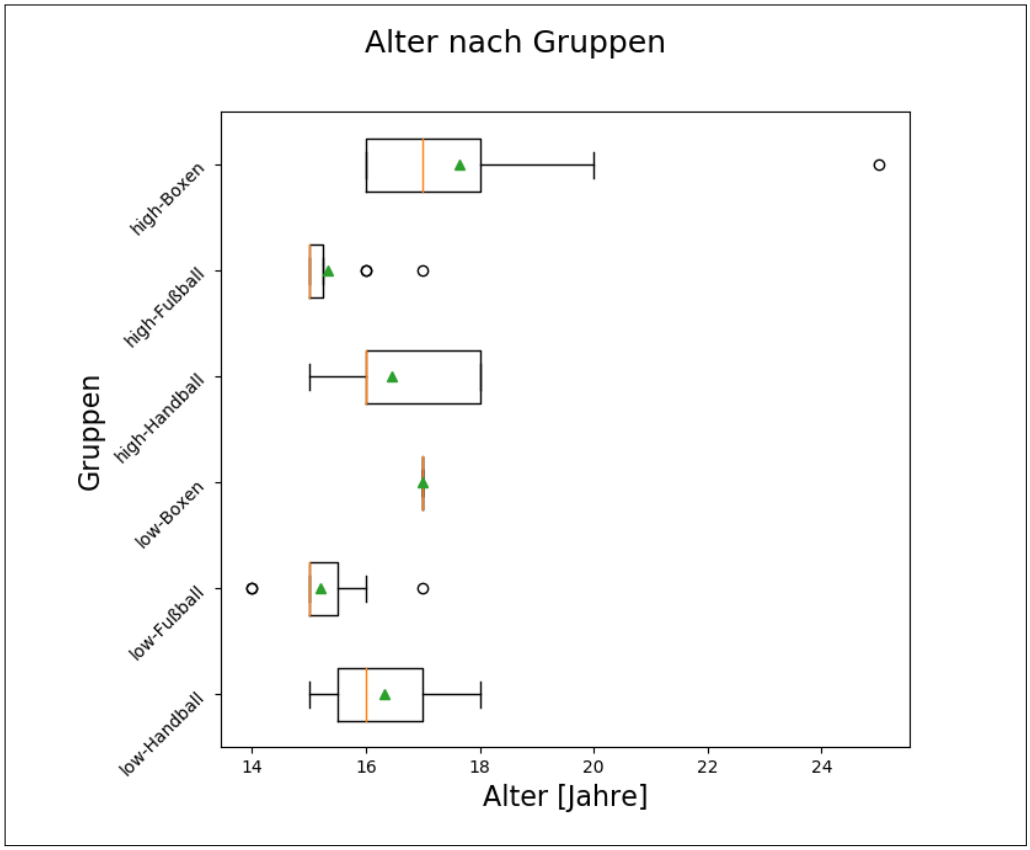

Abbildung 26: Boxplots *Alter nach Gruppen* (Quelle: Eigene Darstellung)

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high-Boxen	16	16	17	17.6 ± 2.3	2	18	20	25
high-Fußball	15	15	15	15.3 ± 0.6	0.2	15.2	15.2	16, 16, 17
high-Handball	15	16	16	16.4 ± 1.2	2	18	18	-
low-Boxen	17	17	17	17 ± 0	0	17	17	-
low-Fußball	15	15	15	15.2 ± 0.7	0.5	15.5	16	14, 14, 17
low-Handball	15	15.5	16	16.3 ± 1.2	1.5	17	18	-

Tabelle 5: Statistische Kenngrößen des Alters [Jahre] (Quelle: Eigene Darstellung)

ren. Das Min beträgt 15 Jahre. Q1 und Med haben einen Wert von 16 Jahren. MW und SD weisen Werte von 16.4 und ±1.2 Jahren auf. Der IQR ist mit 2 Jahren verzeichnet. Q3 und Max betragen 18 Jahre.
Die Gruppe low-Boxen enthält nur ein Individuum mit einem Alter von 17 Jahren.
Der Gruppe low-Fußball sind Werte zwischen 14 und 17 Jahren zugeordnet. Min, Q1 und Med betragen 15 Jahre. MW und SD weisen die Werte 15.2 und ±0.7 Jahre auf. Der IQR ist 0.5 Jahre. Das Q3 besitzt einen Wert von 15.5 und das Max einen Wert von 16 Jahren. Im unteren Bereich existieren zwei Ausreißer von 14 Jahren, im oberen Bereich ein Aus von 17 Jahren.
Die Gruppe low-Handball besitzt einen Wertebereich zwischen 15 und 18 Jahren. Das Min beträgt 15 Jahre, Q1 15.5 und der Med 16 Jahre. MW und SD verzeichnen 16.3 und ±1.2 Jahre. Der IQR liegt bei 1.5 Jahren. Für Q3 sind 17 Jahre angegeben und für das Max 18 Jahre.
Werden die Gruppen bezüglich des Alters miteinander verglichen, so kann festgestellt werden, dass die Individuen der Gruppe high-Boxen mit einem durchschnittlichen Alter von 17.6 Jahren die älteste Gruppe innerhalb der untersuchten Individuen darstellen. Gestützt wird diese Aussage durch die Tatsache, dass 75 % der Individuen zwischen 16 und 18 Jahren alt sind, 25 % sind 18 Jahre alt oder älter. Darüber hinaus verfügt diese Gruppe mit einem Alter von 25 Jahren über das älteste Individuum innerhalb des untersuchten Datensatzes. Des Weiteren weist diese Gruppe die größte Alterspanne auf. Die zweitälteste Gruppe ist low-Boxen. Sie wird gebildet durch ein 17 jähriges Individum. In Bezug auf das Alter kann die Gruppe high-Handball als die Gruppe mit den drittältesten Individuen betrachtet werden. Dies ist zum einen durch einen MW von 16.4 Jahren, einen Med von 16 Jahren sowie die Tatsache zu belegen, dass 75 % der Individuen zwischen 16 und 18 Jahren alt sind. Nur 25 % der Individuen sind 15 Jahre alt. Die Individuen der Gruppe low-Handball sind im Durchschnitt 16.3 Jahre alt. Bei 75 % von ihnen liegt das Alter zwischen 15.5 und 18 Jahren. Somit kann die Gruppe low-Handball als die viertälteste Gruppe angesehen werden. 75 % der Individuen aus den beiden Gruppen high-Fußball und low-Fußball sind zwischen 15 und 15.2 beziehungsweise 15.5 Jahren alt. Auch das durchschnittliche Alter, welches

einen Wert von 15.2 beziehungsweise einen Wert von 15.3 Jahren hat, ist in beiden Gruppen fast identisch. Aus diesen Gründen werden die beiden Gruppen als die jüngsten innerhalb der Betrachtungen angesehen. Des Weiteren enthält die Gruppe low-Fußball mit 14 Jährigen die jüngsten Individuen aller untersuchten Datensätze.

5.2 Einzelbetrachtung der Parameter

In dem vorliegenden Unterkapitel werden die einzelnen Parameter isoliert mit Hilfe von Boxplots für die einzelnen Gruppen statistisch beschrieben und im sportwissenschaftlichen Kontext betrachtet.

5.2.1 Zeit bis zum Abbruch

In Abb. 27 auf S. 93 sind die Boxplots der verschiedenen Gruppen für die tlim zu sehen. Die Tabelle 6 auf S. 92 enthält die zugehörigen statistischen Kenngrößen der Boxplots.
Auf der x-Achse der Abbildung ist die tlim mit einer fünfstufigen Skala von 15 bis 35 min verzeichnet. Die y-Achse enthält die Gruppen[86] *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball*. Med und MW sind innerhalb der Boxplots durch rote vertikale Linien beziehungsweise grüne Dreiecke repräsentiert. Die Tabelle enthält Spalten mit den Überschriften Gruppe, Min, Q1, Med, MW ± SD, IQR, Q3, Max und Aus. Den einzelnen Zeilen können die jeweiligen Kenngrößen der verschiedenen Gruppen entnommen werden.
Die Werte für die Gruppe high-Boxen liegen zwischen 25 (Min) und 35 min (Max). Das Q1 liegt bei 26 min, der Med bei 30 min. MW und SD weisen Werte von 29 beziehungsweise ±3 min auf. Der IQR ist mit 4 min zu verzeichnen. Das Q3 liegt bei 31 min.

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high Boxen	25	26	30	29 ± 3	4	31	35	-
high Fußball	25	25	25	25 ± 0	0	25	25	-
high Handball	25	25	25	26 ± 2	0	25	25	30, 30
low Boxen	20	20	20	20 ± 0	0	20	20	-
low Fußball	15	18	20	19 ± 2	2	20	23	13
low Handball	20	20	20	20 ± 0	0	20	20	-

Tabelle 6: Statistische Kenngrößen der tlim [min] (Quelle: Eigene Darstellung)

[86] Siehe Kapitel 5.1 ab S. 80.

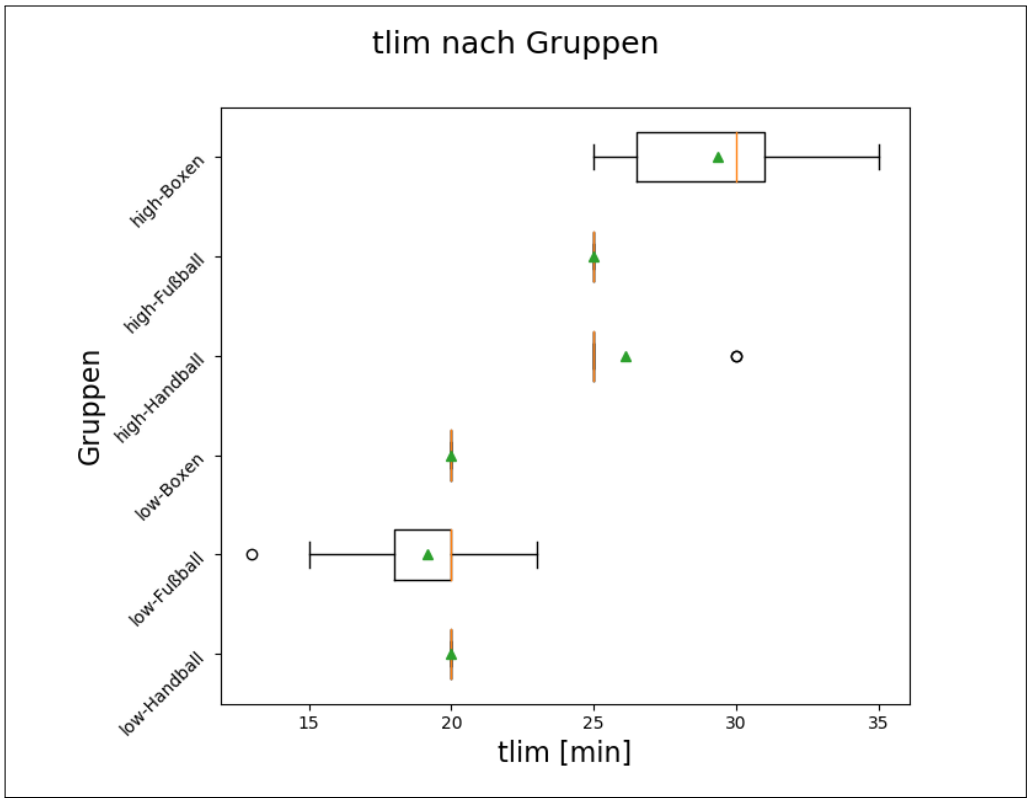

Abbildung 27: Boxplots *tlim nach Gruppen* (Quelle: Eigene Darstellung)

Bei den Kenngrößen der Gruppe high-Fußball sind die Werte der meisten Kenngrößen identisch. So weisen Min, Q1, Med, MW, Q3 und Max einen Wert von 25 min auf. SD und IQR sind 0 min groß. Alle Individuen der Gruppe weisen eine Laufzeit von 25 min auf.
Für die Gruppe high-Handball sind Werte zwischen 25 (Min, Q1, Med, Q3 und Max) und 30 min (Aus) zu verzeichnen. Aufgrund der beiden Aus beträgt der MW 26 min. Die SD weist einen Wert von ±2 min auf, der IQR ist 0 min. Bis auf zwei Ausnahmen existiert für alle weiteren Individuen der Gruppe eine tlim von 25 min.
In der Gruppe low-Boxen gibt es nur ein Individuum mit einer tlim von 20 min.
Die Gruppe low-Fußball besitzt Werte zwischen 13 (Aus) und 23 min (Max). Das Min liegt bei 15 min. Das Q1 hat einen Wert von 18 min. Med und MW betragen 20 beziehungsweise 19 min, SD und IQR ±2 min beziehungsweise 2 min. Das Q3 hat einen Wert von 20 min.
Die Individuen der Gruppe low-Handball weisen alle den gleichen Wert von 20 min auf. SD und IQR sind somit 0 min.
Bei einer vergleichenden Betrachtung zwischen den Clustern, Sportarten und Gruppen sind die folgenden Beobachtungen festzuhalten.
Werden die beiden Cluster high und low miteinander verglichen, liegen die Werte aus Cluster high mit dem niedrigsten Wert von 25 min eindeutig über denen von Cluster low, dessen höchster Wert bei 23 min liegt. Die Sportart Boxen verzeichnet im Mittel mit $\approx$ 29 min die höchsten Werte, gefolgt von Handball mit $\approx$ 25 min. Die im Mittel niedrigsten Werte sind für die Sportart Fußball vorhanden. Der MW in dieser Gruppe beträgt $\approx$ 21 min.
Die Gruppe high-Boxen verzeichnet im Mittel die höchsten Werte sowie die höchste Streuung der Werte für die tlim. Für die Gruppe high-Fußball liegen im Mittel die dritthöchsten Werte vor. Die im Mittel zweithöchsten Werte existieren bei der Gruppe high-Handball. Die beiden Gruppen low-Boxen und low-Handball weisen die vierthöchsten Werte im Mittel auf, gefolgt von der Gruppe low-Fußball, welche im Mittel die niedrigsten Werte aufweist.

Aufgrund des verwendeten Clusterverfahrens[87] ist eine eindeutige Grenze zwischen den Clustern auszumachen. Individuen mit einer tlim von ≥ 25 min sind Cluster high zugeordnet, Individuen mit einem Wert von ≤ 23 min Cluster low. Des Weiteren ist zu beobachten, dass mit Ausnahme der beiden Gruppen high-Boxen und low-Fußball die Individuen der übrigen Gruppen die Belastung nach einer vollständig absolvierten Stufe des Stufentests abgebrochen haben. Diese Tatsache könnte ein Hinweis auf eine mangelnde Motivation der Individuen sein, den Stufentest bis zur absoluten Ausbelastung zu absolvieren. Eine absolute Ausbelastung erscheint im Vergleich bei den Invidiuen der Gruppe high-Boxen und low-Fußball eher gegeben zu sein, da 6 von 14 beziehungsweise 9 von 19 Individuen die Belastung innerhalb einer Stufe abbrachen.
Darüber hinaus kann festgestellt werden, dass die Gruppe high-Boxen die höchsten Werte für die tlim aufweist. Wird davon ausgegangen, dass alle Individuen die maximale Ausbelastung erreicht haben, so sind die Individuen der Gruppe high-Boxen die Individuen mit der höchsten sportlichen Leistungsfähigkeit im Bereich der Ausdauer. Auf diesen Umstand deutet auch die Altersstruktur der Gruppe, welche im Vergleich unter den Gruppen am höchsten liegt. Unter der gleichen Annahme scheinen Individuen der Gruppe low-Fußball mit Werten < 20 min die niedrigste Ausdauerleistungsfähigkeit aufzuweisen. Diese Behauptung würde durch den Umstand gestützt, dass die Individuen dieser Gruppe im Mittel das jüngste Alter im Vergleich zu den Individuen der übrigen Gruppen aufweisen.

Zusammenfassend kann festgestellt werden, dass anhand der tlim ein Vergleich bezüglich der Ausdauerleistungsfähigkeit zwischen verschiedenen Individuen getroffen werden kann. Eine valide Aussage über die Ausdauerleistungsfähigkeit eines einzelnen Individuums kann an dieser Stelle aufgrund der isolierten Betrachtung der tlim jedoch nur bedingt getroffen werden, da bei einer solchen Betrachtung unklar ist, ob die einzelnen Individuen bei Belastungsabbruch ausbelastet waren.

[87] Siehe Kapitel 2.2.1 ab S. 22 und Kapitel 3.1 ab S. 44.

5.2.2 Im Verlauf: Anaerobe Schwelle V4

Abb. 28 auf S. 96 sowie Tabelle 7 auf S. 97 sind die Boxplots und deren statistische Kenngrößen für den Parameter V4 zu entnehmen.
Die Abbildung besitzt auf der x-Achse eine 0.2 stufige Skala mit einem Wertebereich zwischen 3 und 4.2 für die V4 mit der Einheit m/s. Auf der y-Achse sind die Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* sowie *low-Handball* eingezeichnet. Darüber hinaus sind in jedem Boxplot durch eine vertikale rote Linie der jeweilige Med und durch ein grünes Dreieck der jeweilige MW verzeichnet. Die Tabelle enthält Spalten mit den

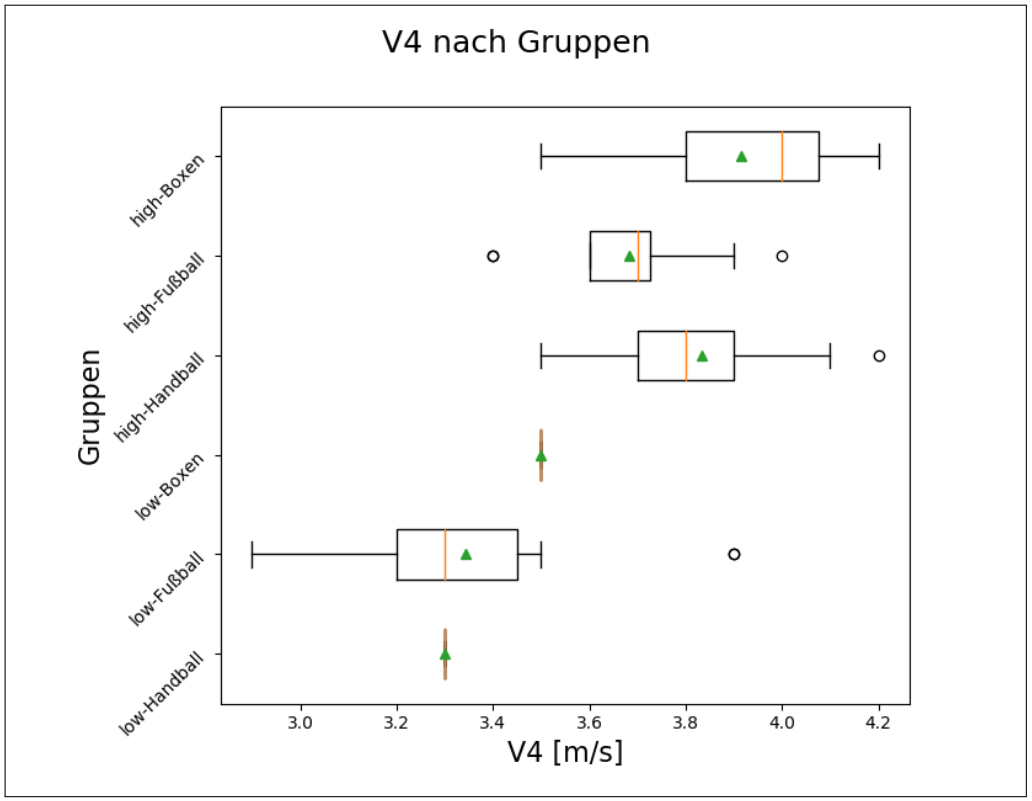

Abbildung 28: Boxplots *V4 nach Gruppen* (Quelle: Eigene Darstellung)

Überschriften Gruppe, Min, Q1, Med, MW ± SD, IQR, Q3, Max und Aus. Pro Zeile der Tabelle sind die jeweiligen Kenngrößen der entsprechenden Gruppen verzeichnet.
Die Gruppe high-Boxen besitzt Werte zwischen 3.5 (Min) und 4.2 m/s (Max).

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high-Boxen	3.5	3.8	4	3.9 ± 0.2	0.3	4.1	4.2	-
high-Fußball	3.6	3.6	3.7	3.7 ± 0.2	0.1	3.7	3.9	3.4, 3.4, 4
high-Handball	3.5	3.7	3.8	3.8 ± 0.2	0.2	3.9	4.1	4.2
low-Boxen	3.5	3.5	3.5	3.5 ± 0	0	3.5	3.5	-
low-Fußball	2.9	3.2	3.3	3.3 ± 0.3	0.2	3.4	3.5	3.9, 3.9
low-Handball	3.3	3.3	3.3	3.3 ± 0	0	3.3	3.3	-

Tabelle 7: Statistische Kenngrößen der V4 [m/s] (Quelle: Eigene Darstellung)

Das Q1 ist mit 3.8 m/s angegeben. Der Med enthält den Wert 4 m/s, MW und SD den Wert 3.9 beziehungsweise ±0.2 m/s. Der IQR liegt bei 0.3 m/s, das Q3 bei 4.1 m/s.
Die Werte der Gruppe high-Fußball liegen zwischen 3.4 (Aus) und 4 m/s (Aus). Das Min und das Q1 betragen 3.6 m/s, Med und MW weisen einen Wert von 3.7 m/s auf, wobei die SD ±0.2 m/s beträgt. Der IQR liegt bei 0.1 m/s, Q3 bei 3.7 m/s und das Max bei 3.9 m/s.
Für die Gruppe high-Handball liegen Werte zwischen 3.5 (Min) und 4.2 m/s (Aus) vor. Das Q1 ist mit 3.7 m/s angegeben, für Med und MW ist jeweils ein Wert von 3.8 m/s zu verzeichnen. Die SD beträgt 0.2 m/s, der IQR ebenfalls 0.2 m/s. Für das Q3 sowie das Max existieren Werte von 3.9 m/s beziehungsweise 4.1 m/s.
Der Wert für das Individuum aus der Gruppe low-Boxen beträgt 3.5 m/s.
Die Gruppe low-Fußball weist Werte zwischen 2.9 (Min) und 3.9 m/s (Aus) auf. Für das Q1 ist ein Wert von 3.2 m/s zu verzeichnen. Der Wert des Meds wie auch der des MWs beträgt jeweils 3.3 m/s, der Wert der SD ±0.3 m/s. Für den IQR ist ein Wert von 0.2 m/s angegeben. Das Q3 weist 3.4 m/s auf, das Max einen Wert von 3.5 m/s.
Die Werte der drei Individuen aus der Gruppe low-Handball betragen für V4 jeweils 3.3 m/s. Entsprechend weisen die statistischen Maße Min, Q1, Med, MW, Q3 und Max des Boxplots somit ebenfalls den Wert 3.3 m/s auf. SD und IQR sind entsprechend 0.

Bei einem Vergleich der Gruppen nach ihrer Clusterzugehörigkeit kann Fol-

gendes festgehalten werden. Die Werte der Gruppen aus Cluster high liegen bis auf einige Ausnahmen oberhalb derer aus Cluster low. Dies wird auch durch die beiden MW der Cluster deutlich. So liegt der MW für Cluster high bei 3.8 m/s, der von Cluster low bei 3.3 m/s. Werden die Sportarten miteinander verglichen, so kann aufgrund der MWe die Sportart Boxen mit einer Ausprägung von 3.9 m/s als diejenige mit den im Mittel höchsten Werten angesehen werden, gefolgt von der Sportart Handball mit einer Ausprägung von 3.7 m/s. Die geringste Ausprägung ist für die Sportart Fußball zu verzeichnen. Die Individuen aus der Gruppe high-Boxen weisen im Mittel die höchsten Werten für V4 auf, gefolgt von der Gruppe high-Handball und high-Fußball. Das Individuum aus der Gruppe low-Boxen weist im Vergleich der Gruppen den vierthöchsten Wert auf. Die beiden Gruppen low-Fußball und low-Handball besitzen im Mittel die niedrigsten Werte. Innerhalb von Cluster high streuen die Werte der dortigen Gruppen gleich stark. Die höchste Streuung ist für die Gruppe low-Fußball zu verzeichnen.

Je höher der Wert für die V4 ist, desto höher ist während des Stufentests die Belastung, unter der additional auf die anaerobe Energiebereitstellung zugegriffen wird.[88] Die Tatsache, dass die Individuen aus der Sportart Fußball in ihren Clustern jeweils die niedrigsten Werte für die V4 aufweisen, könnte zunächst die Frage aufwerfen, ob dieser Umstand der Sportart zuzuordnen ist. An dieser Stelle ist jedoch zu berücksichtigen, dass die Individuen innerhalb dieser Sportart mit einem Alter von im Mittel $\approx$ 15 Jahren unter dem mittleren Alter der Sportarten Boxen ($\approx$ 18 Jahre) und auch Handball ($\approx$ 16 Jahre) liegen. Eine Selektion der Individuen der verschiedenen Sportarten auf eine Altersklasse von 16 Jahren bringt folgende Einsichten hervor: Die 5 Individuen der Sportart Boxen mit einem Alter von 16 Jahren weisen im Mittel eine V4 von $\approx$ 3.9 m/s auf. Die 5 Individuen aus der Sportart Handball besitzen im Mittel einen Wert von $\approx$ 3.7 m/s. Für die 6 Individuen aus der Sportart Fußball existiert ein Wert von 3.5 m/s für die V4. Da diese Individuen jedoch ausschließlich weiblichen Geschlechts sind, wird die Sportart Fußball hier nicht weiter im Vergleich der Sportarten be-

[88] Siehe auch Kapitel 2.3.2 ab S. 36.

trachtet. Wird die V4 als Maß für die Ausdauerleistungsfähigkeit betrachtet, erscheinen die Individuen aus der Sportart Boxen austrainierter zu sein als die Individuen der Sportart Handball. Insgesamt liegen die Werte für die tlim bei den Individuen aus der Sportart Boxen bis auf eine Ausnahme über den Werten der Individuen aus der Sportart Handball.[89] Somit deuten die hier durchgeführten Betrachtungen auf eine Aussagekraft der V4 bezüglich der Ausdauerleistungsfähigkeit hin.

Abschließend kann an dieser Stelle eine Bedeutung der V4 für eine Aussage über die Ausdauerleistungsfähigkeit vermutet werden.

5.2.3 Relative maximale Sauerstoffaufnahme (peak)

In Abb. 29 auf S. 100 sind die Boxplots der Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball* zu finden. Die jeweiligen statistischen Kenngrößen der Boxplots sind der Tabelle 8 auf S. 99 zu entnehmen. Auf der x-Achse der Abbildung ist die rVO_{2peak} in ml/kg/min auf einer fünfstufigen Skala von 30 bis 65 verzeichnet. Auf der y-Achse sind die einzelnen Gruppen aufgetragen. Innerhalb der Boxplots sind Mediane durch eine vertikale rote Linie, arithmetische Mittelwerte durch ein grünes Dreieck gekennzeichnet. Die Tabelle enthält Spalten mit den Überschriften Gruppe, Min, Q1, Med, MW ± SD, IQR, Q3, Max und Aus. Pro

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high-Boxen	50.6	52.1	54	55.1 ± 5.9	6	58.1	66.8	42.3
high-Fußball	44	48.3	50.6	49.9 ± 5.2	4.1	52.5	58.6	37.5
high-Handball	45.6	52	55.5	53.8 ± 4	4.8	56.8	58.1	-
low-Boxen	48	48	48	48 ± 0	0	48	48	-
low-Fußball	38.4	41.7	45.7	44.9 ± 6.1	7.8	49.4	56.5	27.4
low-Handball	42.9	47.3	51.7	51.3 ± 6.7	8.2	55.6	59.4	-

Tabelle 8: Statistische Kenngrößen der rVO_{2peak} [ml/kg/min] (Quelle: Eigene Darstellung)

[89] Siehe Kapitel 5.2.1 ab S. 92.

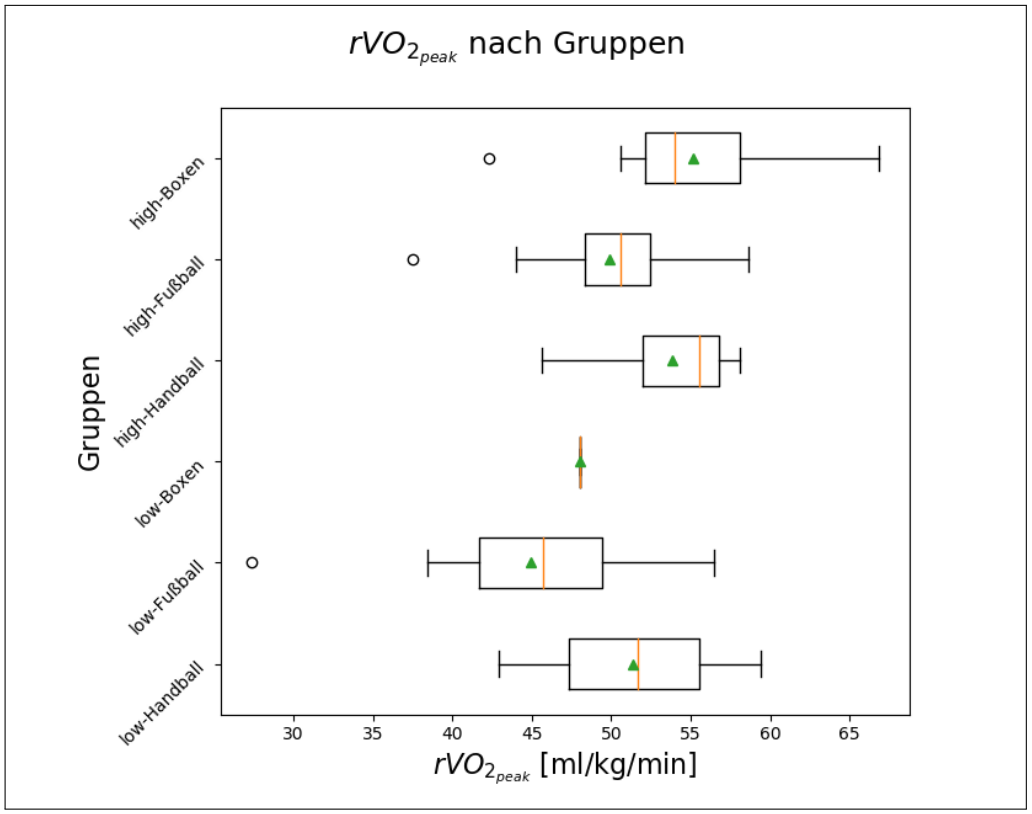

Abbildung 29: Boxplots rVO_{2peak} *nach Gruppen* (Quelle: Eigene Darstellung)

Zeile sind die jeweiligen Kenngrößen der entsprechenden Gruppen verzeichnet.

Als erstes kann der Abbildung der Boxplot für die Gruppe high-Boxen entnommen werden. Dieser umfasst eine Spannweite von 42.3 (Aus) bis 66.8 (Max) ml/kg/min. Das Min beträgt 50.6 ml/kg/min und das Q1 52.1 ml/kg/min. Für den Med sind 54 ml/kg/min zu verzeichnen. Der MW liegt bei 55.1, die SD bei ±5.9 ml/kg/min. Der IQR weist einen Wert von 6 ml/kg/min auf. 58.1 ml/kg/min können für das Q3 entnommen werden.

Die Gruppe high-Fußball besitzt Werte zwischen 37.5 (Aus) und 58.6 (Max) ml/kg/min. Das Min liegt bei 44 ml/kg/min, Q1 bei 48.3 ml/kg/min. Der Med weist einen Wert von 50.6 ml/kg/min auf. MW und SD verzeichnen Werte von 49.9 ml/kg/min beziehungsweise ±5.2 ml/kg/min. Es existiert ein IQR von 4.1 ml/kg/min. Des Weiteren ist ein Wert von 52.5 ml/kg/min für das Q3 zu verzeichnen.

Die Gruppe high-Handball umfasst Werte zwischen 45.6 (Min) und 58.1 (Max) ml/kg/min. Das Q1 liegt bei 52 ml/kg/min. Für den Med ist ein Wert von 55.5 ml/kg/min zu verzeichnen. MW und SD weisen Werte von 53.8 beziehungsweise ±4 ml/kg/min auf, der IQR von 4.8 ml/kg/min. Das Q3 liegt bei 56.8 ml/kg/min.

Das Individuum aus der Gruppe low-Boxen weist einen Wert von 48 ml/kg/min auf.

Die Gruppe low-Fußball besitzt Werte zwischen 27.4 (Aus) und 56.5 (Max) ml/kg/min. Min und Q1 liegen bei 38.4 ml/kg/min beziehungsweise 41.7 ml/kg/min. Der Wert für den Med lautet 45.7 ml/kg/min. MW und SD weisen Werte von 44.9 und ±6.1 ml/kg/min auf. Der IQR beträgt 7.8 ml/kg/min. Für das Q3 und das Max sind Werte von 49.4 ml/kg/min beziehungsweise 56.5 ml/kg/min zu verzeichnen.

Die Werte der Gruppe low-Handball liegen zwischen 42.9 (Min) und 59.4 (Max) ml/kg/min. Q1 und Med liegen bei 47.3 ml/kg/min beziehungsweise 51.7 ml/kg/min. Des Weiteren ist ein MW von 51.3 und eine SD von ±6.7 ml/kg/min zu verzeichnen. Der IQR beträgt 8.2 ml/kg/min. Für das Q3 und das Max betragen die Werte 55.6 und 59.4 ml/kg/min.

Werden die Werte clusterweise verglichen, so kann festgehalten werden, dass die Wertebereiche aus Cluster low Teilmengen der Wertebereiche aus Cluster high sind. Einzige Ausnahme bildet hier der Aus der Gruppe low-Fußball. Die Sportart Boxen weist den zweitgrößten Wertebereich auf, die Sportart Fußball aufgrund des Aus den größten. Der kleinste Wertebereich ist für die Sportart Handball zu verzeichnen. Des Weiteren kann die Sportart Fußball hier als die Sporart mit den niedrigsten Werten im Mittel angesehen werden. Die Gruppe high-Boxen weist im Mittel die höchsten Werte auf. Darüber hinaus liegt bei den Werten dieser Gruppe die höchste Streuung in Cluster high vor. Die Gruppe high-Fußball kann aufgrund ihres MWs als die Gruppe mit den niedrigsten Werten in Cluster high angesehen werden. Darauf deutet auch der Umstand hin, dass das Q3 der Gruppe ungefähr den Wert von Q1 der anderen Gruppen aus Cluster high aufweist. Aufgrund ihres MWs wird die Gruppe high-Handball an dieser Stelle als die Gruppe mit den zweithöchsten Werten betrachtet. Darüber hinaus weist diese Gruppe die geringste Streuung aller Gruppen auf. Die Gruppe low-Boxen besitzt nur ein Individuum. Mit seinem Wert bildet das Individuum die Gruppe mit dem fünfthöchsten Wert im Vergleich der Gruppen. Die Individuen der Gruppe low-Fußball weisen die niedrigsten Werte im Gruppenvergleich auf, was durch den MW zu erkennen ist. Innerhalb von Cluster low weist die Gruppe low-Handball im Mittel die höchsten Werte sowie die höchste Streuung auf.

Werden die Werte der rVO_{2peak} in Bezug auf die tlim betrachtet, können die folgenden Wertekombinationen für die beiden Cluster gefunden und beschrieben werden.[90] Innerhalb von Cluster low können Wertekombinationen mit niedriger rVO_{2peak} und niedriger tlim gefunden werden. Diese treffen auf Individuen der Gruppe low-Fußball zu. Des Weiteren können Kombinationen von einer mittleren rVO_{2peak} und einer niedrigen tlim beobachtet werden. Diese Kombinationen treten hauptsächlich bei Individuen der Gruppe low-Fußball auf, bei zwei Individuen der Gruppe low-Handball und dem Individuum der Gruppe low-Boxen. Eine Wertekombination von hoher

[90] Siehe Kapitel B.1 ab S. 164.

rVO_{2peak} und niedriger tlim findet sich bei je einem Individuum der Gruppe low-Fußball sowie low-Handball.
Für die Individuen des Clusters high sind die folgenden Wertekombinationen zu beobachten. Es existiert eine Wertekombination mit einer niedrigen rVO_{2peak} und einer hohen tlim für ein Individuum aus der Gruppe high-Fußball. Wertekombinationen mit mittleren Werten für die rVO_{2peak} und hohen Werten für die tlim sind in allen Gruppen des Clusters high vertreten. Kombinationen von hohen Werten sowohl bei der rVO_{2peak} als auch bei der tlim sind hauptsächlich bei Individuen aus der Gruppe high-Boxen zu beobachten sowie bei zwei Individuen der Gruppe high-Handball.

Die rVO_{2peak} gilt als eine wichtige Größe für die Leistungsfähigkeit von Atmung, Kreislauf und Muskelstoffwechsel.[91] Diese Annahme wird durch die Werte der Gruppen high-Boxen und high-Handball für die rVO_{2peak} gestützt. Deren Werte sind nämlich höher als die Werte der Gruppen aus Cluster low.
Bei einem Vergleich der Werte aus Gruppe high-Fußball und low-Handball fällt Folgendes auf. Die Individuen aus der Gruppe high-Fußball weisen für die rVO_{2peak} im Mittel niedrigere Werte auf als die Individuen der Gruppe low-Handball. Da die Individuen der Gruppe high-Fußball jedoch ausschließlich eine höhere tlim aufweisen, erscheint die Verwendung der rVO_{2peak} als ausschließliche physiologische Kenngröße zur Bestimmung von Leistungsfähigkeit als nicht sinnvoll. Gestützt wird diese Überlegung durch konkrete Wertekombinationen von der rVO_{2peak} und der tlim, welche im Folgenden beispielhaft aufgeführt werden. So weist ein Individuum der Gruppe high-Boxen eine Wertekombination von 55.2 ml/kg/min rVO_{2peak} sowie eine tlim von 35 min auf. Ein anderes Individium aus der Gruppe low-Fußball weist hingegen bei einer ähnlich hohen rVO_{2peak} von 51.6 ml/kg/min nur eine tlim von 19 min auf.

[91] Siehe auch Kapitel 2.3.2 ab S. 37.

Eine abschließende valide Aussage über den Einfluss der rVO_{2peak} auf die tlim kann an dieser Stelle nicht getroffen werden. Für eine solche ist die Betrachtung weiterer Parameter wie des Parametes RQ_{peak} oder Lak_{peak} notwendig.

5.2.4 Respiratorischer Quotient (peak)

Eine Übersicht über die Werte für den RQ_{peak} der einzelnen Gruppen kann den Boxplots aus Abb. 30 auf S. 105 entnommen werden. Die zugehörigen statistischen Kenngrößen sind darüber hinaus Tabelle 9 auf S. 104 zu entnehmen. Auf der x-Achse der Abbildung ist in einer 0.05 stufigen Skala der RQ_{peak} aufgezeichnet. Die y-Achse enthält die Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball*. Innerhalb der Boxplots sind die jeweiligen Mediane durch vertikale rote Linien, die jeweiligen arithmetischen Mittelwerte durch grüne Dreiecke gekennzeichnet. Die Tabelle enthält Spalten mit den Überschriften Gruppe, Min, Q1, Med, MW ± SD, IQR, Q3, Max und Aus. Pro Tabellenzeile können die Kenngrößen der jeweiligen Gruppen entnommen werden.

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high Boxen	1.03	1.08	1.12	1.11 ± 0.06	0.06	1.14	1.18	1.27
high Fußball	0.97	1.04	1.06	1.06 ± 0.05	0.06	1.1	1.15	-
high Handball	1.01	1.04	1.1	1.1 ± 0.07	0.07	1.11	1.15	1.25
low Boxen	1.06	1.06	1.06	1.06 ± 0	0	1.06	1.06	-
low Fußball	0.99	1.02	1.09	1.07 ± 0.05	0.08	1.1	1.17	-
low Handball	1	1.04	1.09	1.07 ± 0.05	0.06	1.1	1.11	-

Tabelle 9: Statistische Kenngrößen des RQ_{peak} (Quelle: Eigene Darstellung)

Bei Betrachtung der Werte aus der Gruppe high-Boxen ist eine Spannweite zwischen 1.03 (Min) und 1.27 (Aus) zu beobachten. Für das Q1 und den Med können ein Wert von 1.08 beziehungsweise 1.12 entnommen werden. MW und SD sind mit 1.11 und ±0.06 verzeichnet, der IQR mit 0.06. Das Q3 der Gruppe liegt bei 1.14. Darüber hinaus existiert ein Max von 1.18.
Die auftretenden Werte der Gruppe high-Fußball liegen zwischen 0.97 (Min)

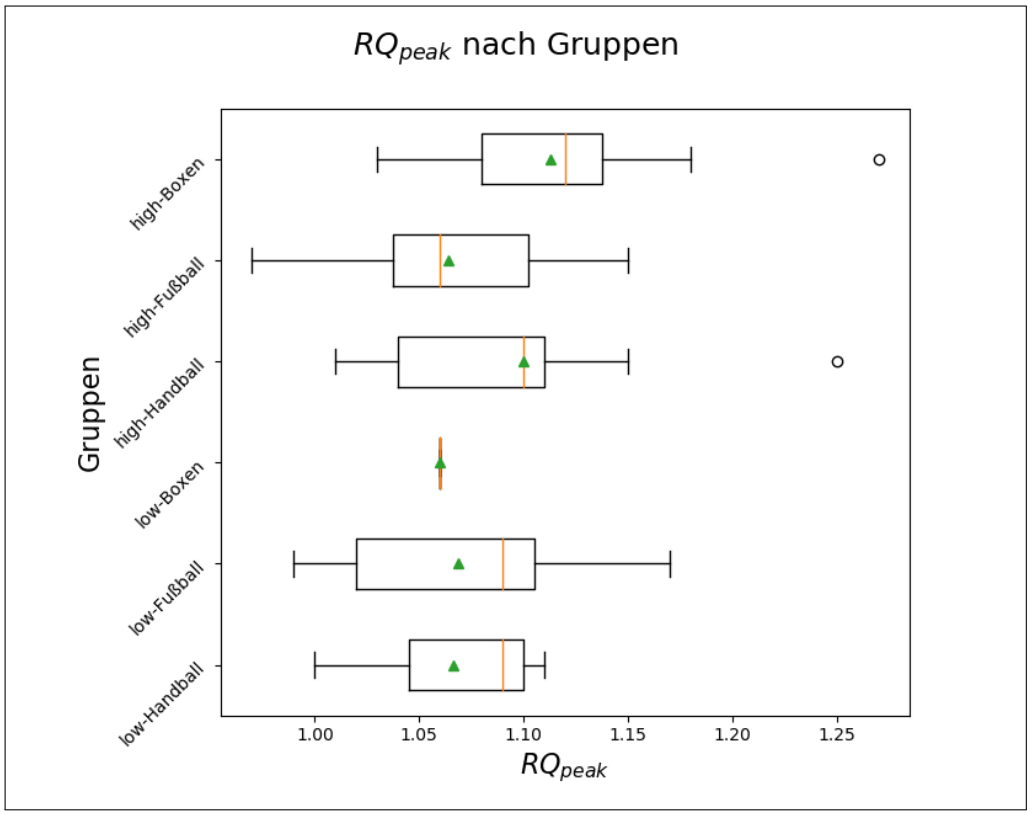

Abbildung 30: Boxplots RQ_{peak} *nach Gruppen* (Quelle: Eigene Darstellung)

und 1.15 (Max), der des Q1 bei 1.04. Med und MW liegen beide bei 1.06, die SD bei ±0.05 und der IQR bei 0.06. Der Wert für das Q3 ist mit 1.1 angegeben.
Für die Gruppe high-Handball können Werte zwischen 1.01 (Min) und 1.25 (Aus) beobachtet werden. Das Q1 weist dabei einen Wert von 1.04, Med und MW beide einen Wert von 1.1 auf. Die SD ist mit ±0.07, der IQR mit 0.07 angegeben. Die Werte für das Q3 und das Max betragen 1.11 beziehungsweise 1.15.
Der Wert des Individuums aus der Gruppe low-Boxen liegt bei 1.06.
Für die Gruppe low-Fußball kann eine Spannweite zwischen 0.99 (Min) und 1.17 (Max) entnommen werden. Das Q1 weist dabei einen Wert von 1.02 auf. Für den Med ist ein Wert von 1.09 verzeichnet, für MW und SD ein Wert von 1.07 beziehungsweise ±0.05. Der IQR ist mit 0.08 angegeben, das Q3 mit 1.1.
Die Werte der Gruppe low-Handball umfassen eine Spannweite zwischen 1 (Min) und 1.11 (Max). Dabei liegt der Q1 bei einem Wert von 1.04 und der Med bei 1.09. MW und SD haben einen Wert von 1.07 beziehungsweise ±0.05. Der Wert für den IQR ist mit 0.06 angegeben, der von Q3 mit 1.1.

Bei einem Vergleich zwischen den beiden Clustern kann festgestellt werden, dass die Spannweiten der Gruppen aus Cluster low eine Teilmenge der Spannweite derer aus Cluster high bilden. Darüber hinaus liegen die Werte aus Cluster high im Mittel mit 1.09 über denen von Cluster low mit 1.07.
Bei der Betrachtung der einzelnen Sportarten sind keine Auffälligkeiten zu finden.
Werden die Werte der einzelnen Gruppen im Mittel miteinander verglichen, so kann festgehalten werden, dass diese innerhalb eines geschlossenen Intervalls zwischen 1.06 und 1.11 auftreten und somit sehr nah beieinander liegen. Ähnlich verhält es sich auch mit den Streuungen der Werte innerhalb der verschiedenen Gruppen. Diese liegen in einem geschlossenen Intervall zwischen 0.05 und 0.07 und damit ebenfalls sehr nah beieinander.

Bei der Betrachtung des RQ_{peak} in Bezug auf die tlim kann Folgendes

festgehalten werden. In einem Bereich zwischen 0.95 und 1.2 können bei verschiedenen Werten des Parameters RQ_{peak} jeweils unterschiedliche Werteausprägungen für die tlim gefunden werden.[92] So existieren beispielsweise Kombinationen mit einem RQ_{peak} zwischen 1.05 und 1.15 und Werte für die tlim zwischen 15 und 35 min. Darüber hinaus fallen zwei sehr hohe Werte von über 1.2 für den RQ_{peak} auf, welche je einem Individuum aus der Gruppe high-Boxen und high-Handball zugeordnet werden können.

Zunächst kann festgestellt werden, dass zum Zeitpunkt des Abbruchs bei keinem Individuum eine reine Fettoxidation, sondern teilweise sogar eine reine Kohlenhydratverbrennung zur Energiebereitstellung vorgelegen hat, da die Ausprägungen für den RQ_{peak} ausschließlich über 0.95 und damit im Bereich von 1 liegen.[93]
55 von 58 Individuen waren nach Tomasits und Haber (2016) beim Belastungsabbruch metabolisch ausbelastet. Dabei waren die Individuen aus Cluster high im Mittel tendenziell etwas höher ausbelastet. Eine höhere metabolische Ausbelastung könnte zum einen mit einem höheren Anteil an laktazidanaerober Energiebereitstellung, zum anderen mit einer höheren Motivation zur sportlichen Leistungserbringung zusammenhängen. Hervorzuheben sind in diesem Kontext besonders die beiden Individuen mit einem sehr hohen Wert von > 1.2 für den RQ_{peak}.
Darüber hinaus kann aufgrund der unterschiedlichen Ausprägungen für die tlim bei gleichen Werten des Parameters RQ_{peak} auf den folgenden Umstand geschlossen werden. Der RQ_{peak} lässt keine Aussage über die Höhe der tlim zu. Vielmehr kann hier festgehalten werden, dass bei einer metabolischen Ausbelastung eines Individuums sehr unterschiedliche Werte bei der tlim erreicht werden können. Dieser Umstand weist auf eine unterschiedliche Ausdauerleistungsfähigkeit der Individuen hin.

Die Frage nach dem Einfluss des RQ_{peak} auf die tlim und damit die sportliche Ausdauerleistungsfähigkeit kann an dieser Stelle nicht valide beantwortet

[92] Siehe Kapitel B.2 ab S. 165.
[93] Siehe auch Kapitel 2.3.2 ab S. 37.

werden. Auch ein Bezug auf die laktazid-anaerobe Energiebereitstellung ist hier ohne die zusätzliche Betrachtung weiterer Parameter wie beispielsweise $\mathrm{Lak}_{\mathrm{peak}}$ nicht ohne Weiteres möglich.

5.2.5 Maximale Herzfrequenz

In Abb. 31 auf S. 109 sind die Boxplots der einzelnen Gruppen für den Parameter $\mathrm{Hf}_{\mathrm{max}}$ zu sehen. Die Tabelle 10 auf S. 108 enthält die statistischen Kenngrößen der Boxplots. In der Abbildung ist auf der x-Achse die $\mathrm{Hf}_{\mathrm{max}}$ mit der Einheit S/min verzeichnet. Die zugehörige Skala ist zehnstufig und reicht von 150 bis 210 S/min. Auf der y-Achse sind die Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball* eingetragen. Der Median und der arithmetische Mittelwert können den jeweiligen Boxplots durch eine rote vertikale Linie beziehungsweise ein grünes Dreieck entnommen werden. Die Tabelle enthält Spalten mit den Überschriften Gruppe, Min, Q1, Med, MW ± SD, IQR, Q3, Max und Aus. Pro Zeile der Tabelle

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high Boxen	186	190	196	198 ± 10	13	203	216	-
high Fußball	184	193	196	196 ± 6	7	200	209	-
high Handball	176	177	182	182 ± 13	13	190	200	153
low Boxen	191	191	191	191 ± 0	0	191	191	-
low Fußball	187	193	198	197 ± 5	6	198	203	208
low Handball	193	197	201	199 ± 4	5	202	203	-

Tabelle 10: Statistische Kenngrößen der $\mathrm{Hf}_{\mathrm{max}}$ [S/min] (Quelle: Eigene Darstellung)

sind die jeweiligen Kenngrößen der entsprechenden Gruppen hinterlegt.
Die Werte der Gruppe high-Boxen liegen zwischen 186 (Min) und 216 (Max) S/min. Das Q1 beträgt 190 S/min, der Med 196 S/min. Der MW und die SD weisen Werte von 198 beziehungsweise ±10 S/min auf. Der IQR beträgt 13 S/min und das Q3 203 S/min.
Für die Gruppe high-Fußball existiert ein Wertebreich zwischen 184 (Min) und 209 (Max) S/min. Das Q1 weist einen Wert von 193 S/min auf, der Med sowie der MW von 196 S/min. Für die SD ist eine Ausprägung von ±6 S/min

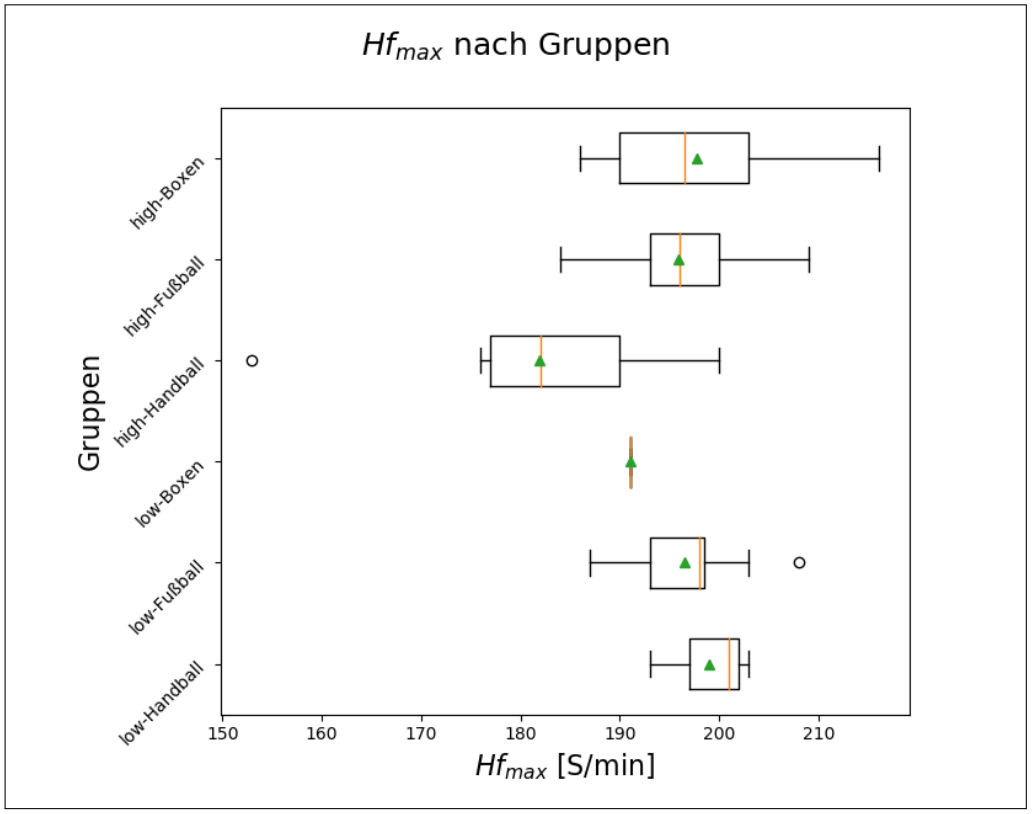

Abbildung 31: Boxplots *Hf_{max} nach Gruppen* (Quelle: Eigene Darstellung)

auszumachen. Der IQR verzeichnet einen Wert von 7 S/min und das Q3 von 200 S/min.
Die Werte der Gruppe high-Handball liegen zwischen 153 (Aus) und 200 (Max) S/min. Für das Min ist ein Wert von 176 S/min festgehalten, für Q1 ein Wert von 177 S/min. Sowohl bei dem Med als auch bei dem MW kann ein Wert von 182 S/min beobachtet werden, wobei die SD ±13 S/min aufweist. Bei einem IQR von 13 S/min beträgt das Q3 190 S/min.
Für das Individuum der Gruppe low-Boxen ist eine Ausprägung von 191 S/min zu verzeichnen.
Der Gruppe low-Fußball können Werte zwischen 187 (Min) und 208 (Aus) S/min zugewiesen werden. Das Q1 liegt dabei bei 193 S/min und der Med bei 198 S/min. MW und SD weisen Werte von 197 beziehungsweise ±5 S/min auf. Der IQR ist 6 S/min, das Q3 198 S/min und das Max 203 S/min.
Für die Gruppe low-Handball existiert ein Wertebereich zwischen 193 (Min) und 203 (Max) S/min. Das Q1 und der Med können mit 197 beziehungsweise 201 S/min bestimmt werden. Für MW und SD ist ein Wert von 199 beziehungsweise ±4 S/min zu entnehmen. Der IQR weist eine Ausprägung von 5 S/min auf, das Q3 von 202 S/min.

Bei einem Vergleich zwischen den Wertebereichen der beiden Cluster kann zunächst festgestellt werden, dass der Wertebereich von Cluster low eine Teilmenge des Wertebereichs von Cluster high ist. Im Mittel liegen die Werte von Cluster high unter denen von Cluster low.[94] Darüber hinaus streuen die Werte von Cluster low geringer als die von Cluster high.
Die Sportart Boxen weist im Mittel die höchsten Werte auf, gefolgt von der Sportart Fußball. Die niedrigsten Werte im Mittel sind für die Sportart Handball zu verzeichnen. Des Weiteren verzeichnet die Sportart Handball die größte Spannweite, die Sportart Boxen die zweitgrößte. Die Sportart Fußball verfügt über die kleinste Spannweite. Mit Ausnahme von high-Handball und low-Boxen liegen die arithmetischen Mittelwerte und Meridiane der verschiedenen Gruppen mit Werten zwischen 196 und 201 S/min sehr nah beieinander. Die im Mittel niedrigsten Werte sind für die Gruppe high-Handball zu

[94] Siehe Kapitel 5.1 ab S. 80.

verzeichnen.

Werden die Werte des Parameters Hf_{max} in Bezug zu der tlim gesetzt, kann Folgendes festgestellt werden.[95] In einem Bereich zwischen 184 und 208 S/min bei der Hf_{max} sind Werte zwischen 13 und 35 min für die tlim zu verzeichnen. In einem Bereich $<$ 184 S/min sind Werte für die tlim von 25 und 30 min zu beobachten. Diese Wertekombinationen sind ausschließlich Individuen der Gruppe high-Handball zuzuordnen. Des Weiteren existieren zwei Wertekombinationen bei Individuen der Gruppe high-Boxen oberhalb von einer Hf_{max} von 210 S/min.

Die breite Streuung von Werten für die tlim in einem Bereich zwischen 184 und 210 S/min bei der Hf_{max} lassen zunächst keinen direkten Einfluss der Hf_{max} auf die tlim vermuten. Als auffällig können jedoch die im Vergleich teilweise niedrigen Werte bei der Hf_{max} angesehen werden, da die entsprechende tlim eine Einordnung in Cluster high ermöglicht. Wird die Betrachtung über einzelne Wertekombinationen hinaus erweitert, so können im Mittel niedrigere Werte bei der Hf_{max} bei Cluster high gegenüber den Werten von Cluster low beobachtet werden. Eine mögliche Erklärung für diese Beobachtung könnte in dem Umstand eines höheren Schlagvolumens des Herzens liegen. Durch ein solches wäre eine höhere tlim bei niedrigerer Herzfrequenz möglich. Neben den hier aufgeführten niedrigen Werten für die Hf_{max} existieren jedoch auch Werte größer als 210 S/min. Diese Beobachtung lässt die Vermutung zu, dass das Schlagvolumen bei diesen Individuen niedriger ausfallen könnte und so das HMV durch eine höhere Herzfrequenz erhöht wird.

Aus den hier aufgeführten Überlegungen kann kein Rückschluss auf den Einfluss der Hf_{max} auf die tlim gezogen werden. Für eine valide Aussage müssen weitere Parameter wie die rVO_{2peak} hinzugezogen werden.

[95] Siehe Kapitel B.3 ab S. 167.

5.2.6 Blutlaktatkonzentration (peak)

Der Abbildung 32 auf S. 113 können die Boxplots der Gruppen aus Cluster high und low für den Parameter Lak_{peak} entnommen werden. Die entsprechenden statistischen Kenngrößen der Gruppen sind Tabelle 11 auf S. 112 zu entnehmen. In der Abbildung ist auf der x-Achse der Parameter Lak_{peak} mit der Einheit mmol/l verzeichnet. Die Skala der Achse ist einstufig und reich von 4 bis 11 mmol/l. Auf der y-Achse sind die Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball* verzeichnet. Med und MW sind in den einzelnen Boxplots durch eine rote vertikale Linie beziehungsweise ein grünes Dreieck gekennzeichnet. Die zugehörige Tabelle enthält Spalten mit den Überschriften Gruppe, Min, Q1, Med, MW ± SD, IQR, Q3, Max und Aus. Pro Zeile sind innerhalb der Tabelle die jeweiligen Kenngrößen der entsprechenden Gruppen verzeichnet.

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high Boxen	5.5	6.4	8.2	8.1 ± 1.9	3.3	9.7	11.1	-
high Fußball	4.3	6.1	6.5	7 ± 1.8	1.8	7.9	9.1	10.7
high Handball	4.6	5.3	5.6	5.8 ± 0.9	0.6	5.9	5.9	7.1, 7.5
low Boxen	4.3	4.3	4.3	4.3 ± 0	0	4.3	4.3	-
low Fußball	4.2	4.7	5.2	5.7 ± 1.5	2.1	6.8	7.3	10.4
low Handball	6.1	6.1	6.1	6.4 ± 0.4	0.4	6.5	6.9	-

Tabelle 11: Statistische Kenngrößen der Lak_{peak} [mmol/l] (Quelle: Eigene Darstellung)

Die Gruppe high-Boxen weist Werte zwischen 5.5 mmol/l (Min) und 11.1 mmol/l (Max) auf, wobei Med und MW 8.2 beziehungsweise 8.1 mmol/l betragen. Neben einer SD von ±1.9 mmol/l beträgt der IQR 3.3 mmol/l. Die Ausprägungen von Q1 und Q3 betragen 6.4 mmol/l beziehungsweise 9.7 mmol/l.
Die Werte der Gruppe high-Fußball erstrecken sich von 4.3 mmol/l (Min) bis 10.7 mmol/l (Aus), Med und MW liegen bei 6.5 mmol/l beziehungsweise 7 mmol/l. Die Werte streuen mit einem Wert von ±1.8 mmol/l (SD). Der IQR weist einen Wert von 1.8 mmol/l auf. Q1 und Q3 weisen 6.1 mmol/l und 7.9

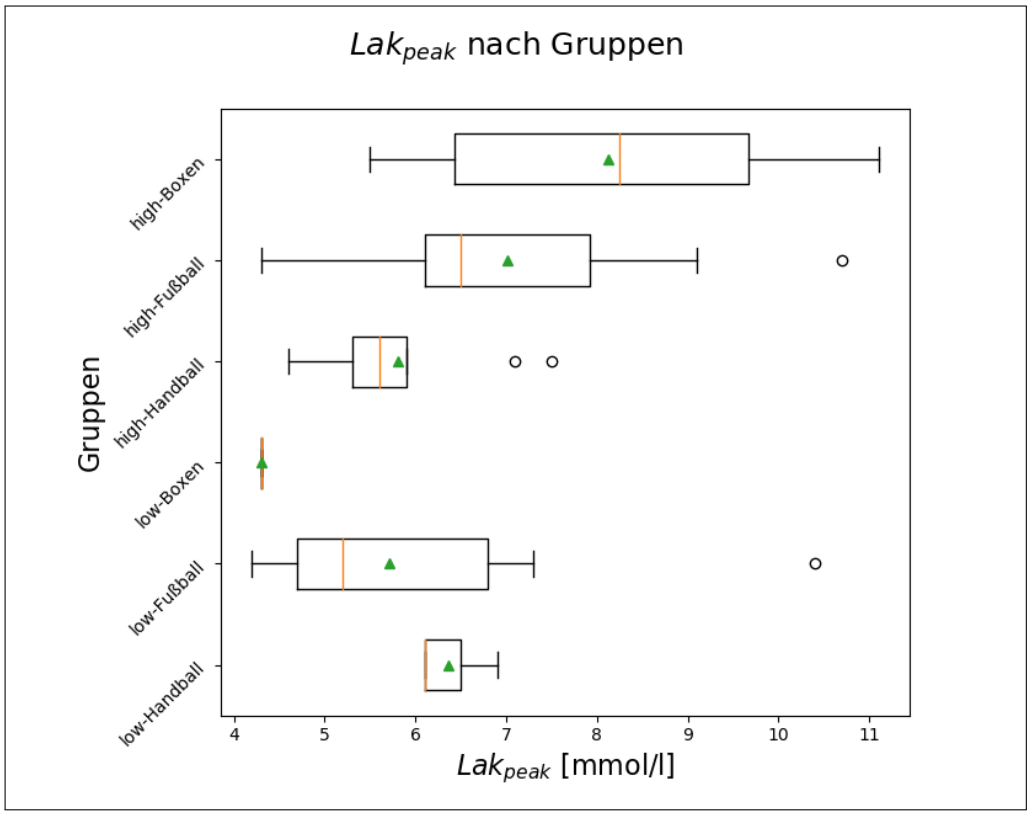

Abbildung 32: Boxplots Lak_{peak} *nach Gruppen* (Quelle: Eigene Darstellung)

mmol/l auf. Das Max beträgt 9.1 mmol/l.
Der Wertebereich für die Gruppe high-Handball liegt zwischen 4.6 mmol/l (Min) und 7.5 mmol/l (Aus). Für Med und MW sind 5.6 mmol/l und 5.8 mmol/l verzeichnet. Des Weiteren streuen die Werte mit ±0.9 mmol/l (SD). Der IQR beträgt 0.6 mmol/l. Für Q1 und Q3 liegen Werte von 5.3 mmol/l beziehungsweise 5.9 mmol/l vor. Der Wert des Max beträgt ebenfalls 5.9 mmol/l. Es existiert ein weiterer Aus von 7.1 mmol/l.
Das Individuum der Gruppe low-Boxen weist einen Wert von 4.3 mmol/l auf.
Die Werte für die Gruppe low-Fußball liegen zwischen 4.2 (Min) und 10.4 mmol/l (Aus). Med und MW weisen Ausprägungen von 5.2 mmol/l beziehungsweise 5.7 mmol/l auf. Die Werte streuen mit ±1.5 mmol/l (SD), wobei der IQR 2.1 mmol/l beträgt. Für Q1 und Q3 liegen Werte von 4.7 mmol/l und 6.8 mmol/l vor. Das Max beläuft sich auf 7.3 mmol/l.
Die Gruppe low-Handball weist Werte zwischen 6.1 (Min, Q1, Med) und 6.9 mmol/l (Max) auf. Der MW liegt bei 6.4 mmol/l wobei die Werte mit ±0.4 mmol/l (SD) streuen. Der IQR beträgt ebenfalls 0.4 mmol/l. Für das Q3 ist eine Ausprägung von 6.5 mmol/l angegeben.

Im Vergleich der Kenngrößen können für die verschiedenen Cluster, Sportarten und Gruppen weitere Beobachtungen aufgeführt werden. Zunächst kann festgehalten werden, dass die Werte aus Cluster low eine Teilmenge des Wertebereichs von Cluster high bilden. Einzige Ausnahme bildet hier das Min der Gruppe low-Fußball. Betrachtet man die einzelnen Sportarten, ist zu erkennen, dass die Sportart Boxen den größten Wertebereich aufweist, gefolgt von der Sportart Fußball. Den kleinsten Wertebereich zeigen die Ausprägungen der Sportart Handball. Somit streuen die Werte der Sportart Handball auch am geringsten. Für die Gruppe high-Boxen liegen im Mittel die höchsten Werte vor. Des Weiteren streuen die Werte dieser Gruppe am höchsten. Von den Werten der Gruppe high-Fußball befinden sich auch Werte unterhalb des Wertebereichs der Gruppe high-Boxen. Alle Werte der Gruppe high-Handball liegen abgesehen von den beiden Ausreißern innerhalb von Q1 der anderen Gruppen aus Cluster high. Darüber hinaus weist diese Gruppe die niedrigste

Streuung innerhalb des Clusters vor. Insgesamt liegen in dieser Gruppe die niedrigsten Werte in Cluster high vor. Innerhalb von Cluster low existiert nur ein Individuum aus der Sportart Boxen. Der Wertebereich der Gruppe low-Fußball ist aufgrund des Ausreißers von 10.4 mmol/l nahezu identisch mit dem der Gruppe high-Fußball. Des Weiteren weist diese Gruppe die höchste Streuung und den höchsten IQR der Gruppen aus Cluster low auf. Die Werte der Gruppe low-Handball weisen die geringste Streuung aller Gruppen auf. Des Weiteren befindet sich der Wertebereich dieser Gruppe über dem Max der Gruppe high-Handball und über dem Wert des Individuums aus der Gruppe low-Boxen. Ansonsten befinden sich die Werte der Gruppe low-Handball innerhalb der Wertebereiche der übrigen Gruppen.

Wird die $\mathrm{Lak}_{\mathrm{peak}}$ in Bezug zur tlim gesetzt, können die im Folgenden aufgeführten Wertekombinationen festgestellt werden.[96] Zunächst kann die Zuordnung von niedrigen Werten für die $\mathrm{Lak}_{\mathrm{peak}}$ zu niedrigen der tlim genannt werden. Hauptsächlich treten diese in der Gruppe low-Fußball, sowie vereinzelt innerhalb der beiden Gruppen low-Boxen und low-Handball auf. Des Weiteren können Kombinationen von mittleren Werten für die $\mathrm{Lak}_{\mathrm{peak}}$ und niedrigen Werten für die tlim ausgemacht werden. Solche Kombinationen sind bei den Gruppen low-Fußball und low-Handball zu finden. In der Gruppe low-Fußball existiert des Weiteren eine Kombination von einem hohen Wert für die $\mathrm{Lak}_{\mathrm{peak}}$ und einer niedrigen tlim.
Für Cluster high können die folgenden Kategorien an Wertekombinationen aufgeführt werden. Für nahezu alle auftretenden Werte der $\mathrm{Lak}_{\mathrm{peak}}$ lässt sich eine Zuordnung zu einer mittleren tlim feststellen. In dieser Kategorie sind Individuen aus allen Gruppen des Clusters high vorzufinden, wobei hier nur die Gruppe high-Fußball bei den hohen Werten für die $\mathrm{Lak}_{\mathrm{peak}}$ vertreten ist. Ferner lässt sich eine niedrigere $\mathrm{Lak}_{\mathrm{peak}}$ in Kombination mit einer hohen tlim ausmachen. Diese Kombinationen sind für Individuen der Gruppen high-Boxen und high-Handball zu erkennen. Als letzte auftretende Kombination kann eine hohe $\mathrm{Lak}_{\mathrm{peak}}$ mit einer hohen tlim aufgeführt werden. Diese Kombination ist nur bei Individuen der Gruppe high-Boxen zu finden.

[96] Siehe Kapitel B.4 ab S. 169.

Die hohen Werte für die Lak_{peak} lassen im Gegensatz zu den niedrigeren Werten zunächst auf einen höheren Zugriff auf die anaerobe Energiebereitstellung schließen. Da es sich bei der Messung jedoch um die Laktatkonzentration im Blut und nicht um die Konzentration in einem Mitochondrium oder in einer aktiven Muskelzelle handelt, kann hier aus folgenden Gründen nicht ausschließlich auf einen höheren Anteil der anaeroben Energiebereitstellung geschlossen werden. Aufgrund der unterschiedlichen Ausprägung von Laktattransport- und Laktateliminationsprozessen könnte auch bei den niedrigeren Werten für die Lak_{peak} ein hoher Anteil an anaerober Energiebereitstellung vorliegen. Dies könnte beispielsweise bei den Individuen mit niedriger Lak_{peak} und hoher tlim der Fall sein. Aufgrund von möglicherweise besser ausgeprägten Laktattransport- und Laktateliminationsprozessen könnte hier die Lak_{peak} niedriger ausfallen.

Eine abschließende valide Aussage über den Einfluss der Lak_{peak} auf die tlim zu treffen, erscheint hier unmöglich. Für eine solche Aussage sind weitere Parameter wie beispielsweise die $\text{rVO}_{2_{\text{peak}}}$ oder der RQ_{peak} in tiefergehende Betrachtungen einzubeziehen.

5.2.7 Hämoglobin-Wert

Der Abbildung 33 auf S. 117 können die Boxplots der Gruppen *high-Boxen*, *high-Fußball*, *high-Handball*, *low-Boxen*, *low-Fußball* und *low-Handball* für den Parameter Hb entnommen werden. Die statistischen Kenngrößen der Boxplots sind darüber hinaus in Tabelle 12 auf S. 118 verzeichnet. Auf der x-Achse ist der physiologische Parameter Hb mit einer Einheit von g/dl eingezeichnet. Die einstufige Skala reicht von 11 bis 16 g/dl. Die y-Achse weist die einzelnen Gruppen auf. Mediane und arithmetische Mittelwerte sind pro Boxplot durch eine rote vertikale Linie beziehungsweise ein grünes Dreieck gekennzeichnet. Die Spalten der Tabelle beinhalten die Überschriften Gruppe, Min, Q1, Med, MW $\pm$ SD, IQR, Q3, Max und Aus. In den einzelnen Zeilen

der Tabelle befinden sich die jeweiligen Kenngrößen der einzelnen Gruppen.

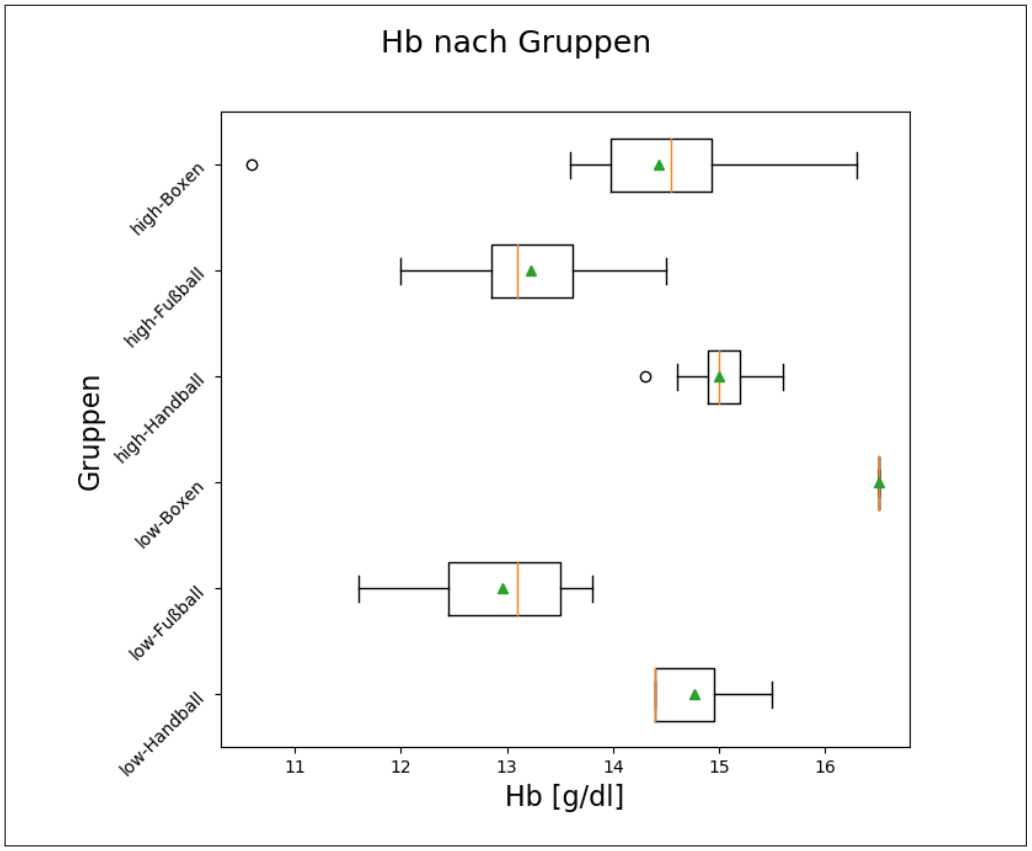

Abbildung 33: Boxplots *Hb nach Gruppen* (Quelle: Eigene Darstellung)

Die Gruppe high-Boxen besitzt einen Boxplot mit einer Spannweite zwischen 10.6 (Aus) und 16.3 (Max) g/dl. Das Min des Boxplots beträgt 13.6 g/dl. Für das Q1 ist ein Wert von 14 g/dl verzeichnet, für den Med eine Ausprägung von 14.6 g/dl. MW und SD beinhalten Werte von 14.4 und ±1.3 g/dl. Der IQR ist mit 0.9 g/dl angegeben, das Q3 mit 14.9 g/dl.
Für die Gruppe high-Fußball liegen Werte zwischen 12 (Min) und 14.5 (Max) g/dl vor. Das Q1 beträgt 12.8 g/dl, der Med 13.1 g/dl. Für den MW und die SD sind die Ausprägungen 13.2 beziehungsweise ±0.7 g/dl zu entnehmen. Der IQR liegt bei 0.8 g/dl, das Q3 bei 13.6 g/dl.
Werte zwischen 14.3 (Aus) und 15.6 (Max) g/dl existieren für die Gruppe high-Handball. Das Min beträgt dabei 14.6 g/dl, das Q1 14.9 g/dl. Für den

Gruppe	Min	Q1	Med	MW ± SD	IQR	Q3	Max	Aus
high Boxen	13.6	14	14.6	14.4 ± 1.3	0.9	14.9	16.3	10.6
high Fußball	12	12.8	13.1	13.2 ± 0.7	0.8	13.6	14.5	-
high Handball	14.6	14.9	15	15 ± 0.4	0.3	15.2	15.6	14.3
low Boxen	16.5	16.5	16.5	16.5 ± 0	16.5	16.5	16.5	-
low Fußball	11.6	12.4	13.1	13 ± 0.6	1.1	13.5	13.8	-
low Handball	14.4	14.4	14.4	14.8 ± 0.5	0.5	15	15.5	-

Tabelle 12: Statistische Kenngrößen des Hb [g/dl] (Quelle: Eigene Darstellung)

Med ist eine Ausprägung von 15 g/dl zu verzeichnen. Ebenso beträgt der MW 15 g/dl, die SD lautet ±0.4 g/dl. Der IQR ist mit 0.3 g/dl angegeben, das Q3 mit 15.2 g/dl.
Der Wert für das Individuum der Gruppe low-Boxen lautet 16.5 g/dl.
Der Wertebereich der Gruppe low-Fußball liegt zwischen 11.6 (Min) und 13.8 (Max) g/dl. Das Q1 ist mit 12.4 g/dl angegeben, der Med mit 13.1 g/dl. Für MW und SD sind 13 beziehungsweise ±0.6 g/dl zu verzeichnen. Der IQR beträgt 1.1 g/dl. Das Q3 liegt bei 13.5 g/dl.
Der Gruppe low-Handball können Werte zwischen 14.4 (Min) und 15.5 (Max) g/dl zugeordnet werden. Sowohl das Min, das Q1 als auch der Med weisen alle eine Ausprägung von 14.4 g/dl auf. MW und SD betragen 14.8 beziehungsweise ±0.5 g/dl. Q3 liegt bei 15 g/dl.

Beim Vergleich der Werte zwischen den Clustern kann festgestellt werden, dass der Wertebereich aus Cluster low eine Teilmenge des Wertebereichs von Cluster high ist. Eine Ausnahme bildet hier der Wert der Gruppe low-Boxen. Die Sportart Boxen weist den höchsten Wertebereich auf, gefolgt von der Sportart Fußball. Die Sportart Handball verfügt über den kleinsten Wertebereich. Darüber hinaus kann die Sportart Handball als die mit den im Mittel höchsten Werten (14.9 g/dl) angesehen werden, gefolgt von Boxen (14.6 g/dl) und Fußball (13.1 g/dl).
Die Gruppe high-Boxen weist im Mittel die vierthöchsten Werte und die höchste Streuung auf. Des Weiteren ist in der Gruppe der niedrigste Wert al-

ler betrachteten Datensätze vorhanden. Die Werte der Gruppe high-Fußball besitzen im Mittel die zweitniedrigsten Ausprägungen. Die Gruppe high-Handball besitzt im Mittel die zweithöchsten Werte und die geringste Streuung. Die Gruppe low-Boxen weist den höchsten Wert aller Datensätze auf. Die Werte der Gruppe low-Fußball können im Mittel als die niedrigsten beobachtet werden. Alle Werte dieser Gruppe liegen unterhalb der Werte der Gruppen high- und low-Boxen. Die Gruppe low-Handball weist die dritthöchsten Werte im Mittel sowie die zweitniedrigste Streuung auf.

Wird der Hb in Bezug zur tlim gesetzt, können dabei verschiedene Wertekombinationen beobachtet werden.[97] Zunächst sind Wertekombinationen mit niedrigem Hb und niedriger tlim zu finden. Diese Kombinationen sind ausschließlich Individuen der Gruppe low-Fußball zuzuordnen. Als eine weitere Kombination können hohe Werte für Hb mit niedrigen Werten für die tlim genannt werden. Diese Kombination tritt für einige wenige Individuen aller Gruppen aus Cluster low auf. Des Weiteren ist die Kombination von niedrigen Werten für Hb sowie hohen Werten für die tlim zu finden. Diese Kombination ist hauptsächlich durch die Gruppe high-Fußball vertreten. Einzige Ausnahme bildet ein Individuum aus der Gruppe high-Boxen. Wertekombinationen mit hohen Ausprägungen von Hb und der tlim können für alle Individuen aus Cluster high gefunden werden.

Indem es Sauerstoff (O_2) bindet, besitzt Hämoglobin eine Funktion beim Transport von O_2 und damit beim aeroben Metabolismus.[98] Bei einem Vergleich der arithmetischen Mittelwerte aus den verschiedenen Gruppen kann zunächst festgehalten werden, dass die Mittelwerte der Gruppen aus Cluster low teilweise höher sind als die der Gruppen aus Cluster high. Auffällig ist, dass die Gruppen high- und low-Fußball im Mittel die niedrigsten Werte aufweisen. Dieser Umstand liegt mit hoher Wahrscheinlichkeit in der Tatsache begründet, dass diese Gruppen nur aus weiblichen Individuen bestehen und solche niedrigere Referenzwerte für den Hb aufweisen, als dies bei männlichen

[97] Siehe Kapitel B.5 ab S. 171.
[98] Siehe auch Kapitel 2.3.2 ab S. 39.

Individuen der Fall ist.
Des Weiteren kann Folgendes festgestellt werden. Ein hoher Wert für den Parameter Hb lässt nicht auf einen hohen Wert für die tlim schließen.[99] So existieren Wertekombinationen der Parameter Hb und tlim von über 16 g/dl für den Hb und einer tlim von 20 min, 25 min und sogar 30 min. Des Weiteren existieren Werte für den Hb zwischen 12 und 13 g/dl bei einer tlim von 18 min, 20 min aber auch von 25 min.
Darüber hinaus kann an dieser Stelle angemerkt werden, dass für ein Individuum der Gruppe high-Boxen ein Hb von 10.6 g/dl existiert. Damit liegt dieser Wert außerhalb der Referenzwerte für männliche Individuen. Ob es sich an dieser Stelle um einen möglichen Messfehler, einen Fehler bei der Datenintegration oder -verarbeitung oder aber einen pathologischen Wert handelt, kann an dieser Stelle nicht festgestellt werden.

Abschließend kann bei der Betrachtung des Parameters Hb an dieser Stelle Folgendes festgehalten werden. Die isolierte Betrachtung dieses Parameters lässt keinen direkten Einfluss auf den Parameter tlim erkennen. Somit kann hier eine Aussage bezüglich eines Einflusses auf die sportliche Leistungsfähigkeit eines Individuums nicht getroffen werden. Für eine solche sind weitere Parameter in die Betrachtung mit einzubeziehen.

[99] Siehe Kapitel B.5 ab S. 171.

5.3 Parameterübergreifende Regeln

Aus Abb. 34 auf S. 121 kann der Decision Tree[100] für die physiologischen Parameter der Cluster entnommen werden. Der Entscheidungsbaum ent-

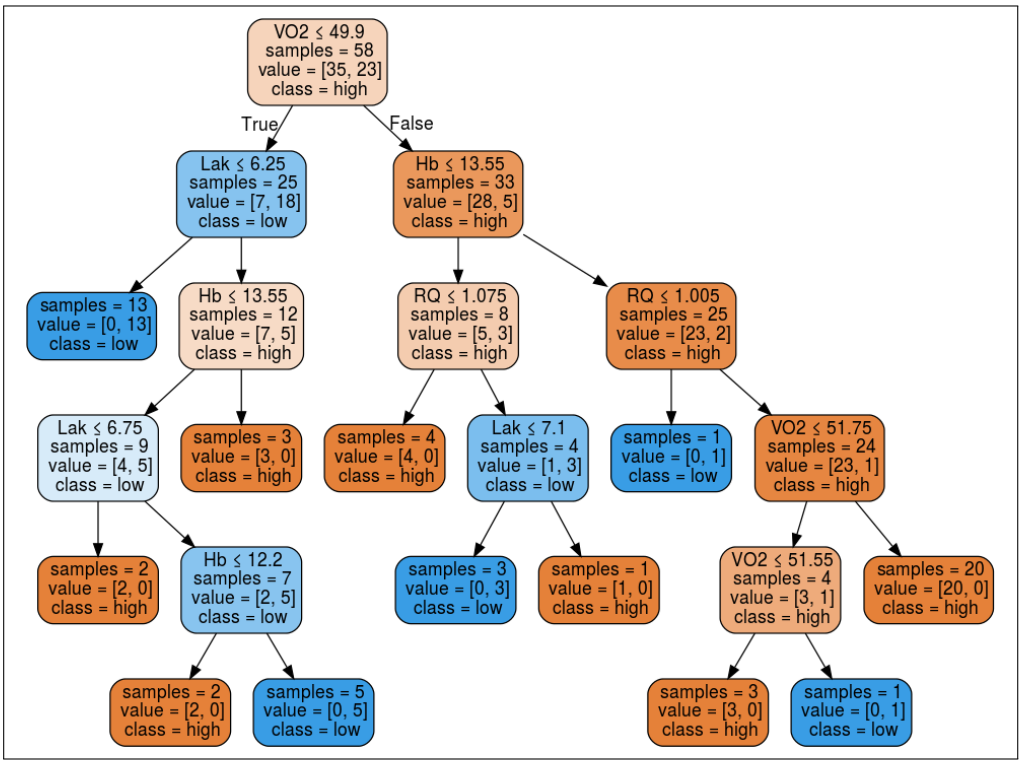

Abbildung 34: Decision Tree der physiologischen Parameter (Quelle: Eigene Darstellung)

hält Knoten, die die physiologischen Parameter $\mathrm{rVO2_{peak}}$ (Knoten *VO2*), $\mathrm{Lak_{peak}}$ (Knoten *Lak*), Hb (Knoten *Hb*) sowie $\mathrm{RQ_{peak}}$ (Knoten *RQ*) enthalten. Der Parameter $\mathrm{Hf_{max}}$ hat für die Einordnung eines Individuums in ein Cluster innerhalb dieses Modells keine Bedeutung. Knoten mit einer Mehrzahl an Individuen aus Cluster high sind entsprechend hellorange bis dunkelorange gefärbt, Knoten mit einer Mehrzahl an Individuen aus Cluster low hellblau bis dunkelblau. Je dunkler ein Knoten ist, auf desto mehr Individuen eines entsprechenden Clusters wird die Bedingung des Knotens verhältnismäßig angewendet. Ein Knoten enthält zunächst eine Bedingung

[100] Siehe auch Kapitel 2.2.2 ab S. 27.

und verschiedene Attribute. Des Weiteren kann einem Knoten die Anzahl an Individuen entnommen werden, für welche die Bedingung zutrifft (*samples*). Wie viele Individuen von den samples den Klassen high oder low zuzuordnen sind, ist dem Attribut *value* zu entnehmen. Schlussendlich kann die Klasse selbst entnommen werden (*class*). Diese richtet sich nach der größeren Anzahl an Individuen innerhalb des Knotens. Für Kanten des Baums, welche auf der linken Seite des Baums angeordnet sind, gilt True, für Kanten auf der rechten Seite False. Die Beschreibung des Baums wird ebenenweise – beginnend mit dem Wurzelknoten und endend mit den Blättern des Baumes – von links nach rechts durchgeführt.
Der Wurzelknoten enthält die Bedingung $VO2 \leq 49.9$, welche auf alle 58 Individuen – 35 Individuen aus Cluster high und 23 Individuen aus Cluster low – angewendet wird. Die Klasse des Knotens ist high.
In der 1. Ebene unterhalb des Wurzelknotens existieren zwei innere Knoten. Der erste Knoten der 1. Ebene enthält die Bedingung $Lak \leq 6.25$. Diese Bedingung wird auf 25 Individuen, 7 aus Cluster high und 18 aus Cluster low angewendet. Der Knoten selbst ist der Klasse low zugeordnet. Der zweite Knoten der 1. Ebene weist die Bedingung $Hb \leq 13.55$ auf. Der Knoten enthält 33 Individuen, von denen 28 Cluster high und 5 Individuen Cluster low zugeordnet sind. Die Klasse des Knotens ist high.
Auf der Ebene 2 sind ein Blatt und drei innere Knoten enthalten. Das Blatt gilt für 13 Individuen aus Cluster low. Entsprechend bildet low die Klasse des Blattes. Der erste innere Knoten der 2. Ebene enthält ebenfalls die Bedingung $Hb \leq 13.55$, die durch den Knoten auf 7 Individuen aus Cluster high, 5 aus Cluster low und damit auf insgesamt 12 Individuen angewendet wird. Da die Anzahl an Individuen aus Cluster high überwiegt, ist der Knoten entsprechend mit der Klasse high bezeichnet. Der zweite innere Knoten weist die Bedingung $RQ \leq 1.075$ auf. Diese gilt für 8 Individuen, von denen 5 aus Cluster high und 3 aus Cluster low stammen. Die Klasse des Knotens lautet high. Der dritte innere Knoten auf Ebene 2 besitzt die Bedingung $RQ \leq 1.005$. Diese Bedingung wird für 23 Individuen aus Cluster high, für 2 aus Cluster low angewendet und ist somit für insgesamt 25 Individuen erfüllt. Dadurch ergibt sich für den Knoten die Klasse high.

In der 3. Ebene existieren drei innere Knoten und drei Blätter. Der erste innere Knoten auf Ebene 3 weist die Bedingung $Lak \leq 6.75$ auf. Diese findet Anwendung auf 9 Individuen, 4 aus Cluster high und 5 aus Cluster low und ist der Klasse low zugeordnet. Der zweite innere Knoten auf der 3. Ebene beinhaltet die Bedingung $Lak \leq 7.1$, die für 4 Individuen gilt. 1 Individuum ist dabei Cluster high zugeordnet, 3 Individuen Cluster low. Aufgrund dessen ist die Klasse low für den Knoten verzeichnet. Der dritte innere Knoten auf Ebene 3 weist die Bedingung $VO2 \leq 51.75$ auf. Diese gilt für 24 Individuen. Bis auf 1 Individuum sind alle Individuen innerhalb dieses Knotens Cluster high zugeordnet. Die Klasse des Knotens ist dadurch high. Die ersten beiden Blätter der Ebene gelten für 3 beziehungsweise 4 Individuen aus Cluster high (class = high). Das dritte Blatt gilt für 1 Individuum aus Cluster low (class = low).
Ebene 4 beinhaltet zwei innere Knoten und vier Blätter. Der erste innere Knoten mit der Bedingung $Hb \leq 12.2$ gilt für 7 Individuen, 2 aus Cluster high und 5 aus Cluster low. Die Klasse des Knotens ist daher low. Die Bedingung $VO2 \leq 51.55$ ist im zweiten Knoten von Ebene 3 zu finden. Der Knoten beinhaltet 3 Individuen aus Cluster high und 1 Individuum aus Cluster low. Somit ist der Knoten der Klasse high zugeordnet. Die Blätter eins, drei und vier gelten für 2, 1 und 20 Individuen aus Cluster high und sind somit Klasse high zugeteilt. Das zweite Blatt der Ebene ist der Klasse low zugeordnet und gilt für 3 Individuen aus Cluster low.
Auf der 5. und letzten Ebene befinden sich ausschließlich Blätter. Blatt eins und drei gelten für 2 beziehungsweise 3 Individuen aus Cluster high und gehören dementsprechend der Klasse high an. Die beiden Blätter zwei und vier sind der Klasse low zugeordnet und gelten für 5 Individuen beziehungsweise 1 Individuum aus Cluster low. Werden die Pfade innerhalb des Decision Trees vom Wurzelknoten bis zu den Blättern durchlaufen, können zwölf Regeln abgeleitet werden. Diese Regeln sind Tabelle 13 auf S. 124 zu entnehmen. Die Tabelle enthält die Spaltenüberschriften Cluster (Clus), Regel, $\mathrm{rVO2_{peak}}$ [ml/kg/min], $\mathrm{RQ_{peak}}$, $\mathrm{Lak_{peak}}$ [mmol/l], Hb [g/dl] und Anzahl (Anz). Jeder Spalte ist eine Regel zu entnehmen. Die Reihenfolge der Regeln innerhalb der Tabelle basiert auf der jeweiligen Individuenanzahl. Regeln mit

Clus	Regel	$rVO2_{peak}$ [ml/kg/min]	RQ_{peak}	Lak_{peak} [mmol/l]	Hb [g/dl]	Anz
high	1	> 51.75	> 1.005		> 13.55	20
	2	> 49.9	≤ 1.075		≤ 13.55	4
	3	≤ 49.9		> 6.25	> 13.55	3
	4	(49.9, 51.55]	> 1.005		> 13.55	3
	5	≤ 49.9		(6.25, 6.75]	≤ 13.55	2
	6	≤ 49.9		> 6.75	≤ 12.2	2
	7	> 49.9	> 1.075	> 7.1	≤ 13.55	1
low	8	≤ 49.9		≤ 6.25		13
	9	≤ 49.9		> 6.75	[12.2, 13.55]	5
	10	> 49.9	> 1.075	≤ 7.1	≤ 13.55	3
	11	> 49.9	≤ 1.005		> 13.55	1
	12	(51.55, 51.75]	> 1.005		> 13.55	1

Tabelle 13: Regeln für die physiologischen Parameter

einer höheren Individuenanzahl werden weiter oben in der Tabelle aufgelistet als solche mit einer geringeren Individuenanzahl. Ist die Individuenanzahl zweier Regeln gleich, so wird die Regel zuerst aufgeführt, die dem Decision Tree von links nach rechts gelesen als erstes zu entnehmen ist.
Die ersten sieben Regeln können Individuen aus Cluster high zugeordnet werden. Die restlichen fünf Regeln entstammen Parameterausprägungen von Individuen aus Cluster low. Im weiteren Verlauf werden nur die Regeln betrachtet, die auf mindestens 3 Individuen zutreffen.
Regel 1 besagt, dass die $rVO2_{peak}$ größer als 51.75 ml/kg/min, der RQ_{peak} größer als 1.005 und der Hb größer als 13.55 g/dl ist. Die Regel gilt für insgesamt 20 Individuen aus Cluster high und damit für ≈ 57 % aller Individuen aus Cluster high. Diese 20 Individuen entstammen zu 11 Individuen der Gruppe high-Boxen, zu 2 Individuen der Gruppe high-Fußball und zu 7 Individuen der Gruppe high-Handball. Sie weisen ein Alter von 15, 16, 17, 18, 20 und 25 Jahren auf. Somit ist die Regel für die vorliegenden Datensätze weder spezifisch in Bezug auf Sportart noch Alter. Bezogen auf die tlim kann folgende Aussage durch die Regel getroffen werden. Um eine tlim von

mindestens 25 min erzielen zu können, müssen für die Parameter $rVO_{2_{peak}}$, RQ_{peak} und Hb mindestens die Ausprägungen 51.75 ml/kg/min, 1.005 beziehungsweise 13.55 g/dl vorhanden sein.
Nach Regel 2 müssen für die $rVO_{2_{peak}}$ Werte über 49.9 ml/kg/min vorliegen. Des Weiteren dürfen die Werte für den Parameter RQ_{peak} höchstens 1.075 und für den Hb höchstens 13.55 g/dl betragen. Diese Regel findet auf 4 Individuen aus der Gruppe high-Fußball mit einer Altersstruktur von 15 und 17 Jahren Anwendung. Bei den Werten des Parameters RQ_{peak} ist zu erkennen, dass es sich bei den Grenzen um berechnete Werte handelt. Diese entsprechen damit nicht gemessenen Werten des Parameters. In Bezug auf die tlim kann die folgende Aussage für Regel 2 getroffen werden. Um eine tlim von 25 min erbringen zu können, muss der Parameter $rVO_{2_{peak}}$ einen Wert größer als 49.9 ml/kg/min aufweisen. Die Parameter RQ_{peak} und Hb überschreiten dabei nicht die Ausprägungen von 1.075 beziehungsweise 13.55 g/dl.
Für Regel 3 gilt, dass die $rVO_{2_{peak}}$ höchstens 49.9 ml/kg/min, die Lak_{peak} größer als 6.25 mmol/l und der Hb größer als 13.55 g/dl sein müssen. Die Regel ist für 1 Individuum aus der Gruppe high-Fußball und 2 Individuen aus der Gruppe high-Handball gültig. Die Altersstruktur beträgt 15, 16 und 18 Jahre. Verknüpft mit der tlim kann die folgende Aussage getroffen werden. Um eine tlim von 25 min erbringen zu können, ist eine $rVO_{2_{peak}}$ von höchstens 49.9 ml/kg/min ausreichend, wenn die Werte der Parameter Lak_{peak} und Hb über 6.25 mmol/l beziehungsweise über 13.55 g/dl liegen.
Regel 4 enthält die folgenden Bedingungen. Die $rVO_{2_{peak}}$ muss größer als 49.9 ml/min/kg sein und darf höchstens 51.55 ml/kg/min betragen. Des Weiteren müssen der RQ_{peak} größer als 1.005 und der Hb größer als 13.55 g/dl sein. Die Regel trifft auf ein 1 Individuum aus der Gruppe high-Fußball zu, das ein Alter von 15 Jahren besitzt. Ebenso gilt die Regel für 2 Individuen aus der Gruppe high-Boxen mit einem Alter von 16 beziehungsweise 17 Jahren. In Bezug auf die tlim kann die folgende Aussage für Regel 4 getroffen werden. Eine tlim von 25 min kann mit einer $rVO_{2_{peak}}$ zwischen 49.9

und einschließlich 51.55 ml/kg/min erbracht werden. Der RQ_{peak} muss dafür über 1.005 liegen und der Hb über 13.55 g/dl.
Regeln 5 bis 7 werden nicht weiter betrachtet, da sie nur auf die Werte von 1 oder maximal 2 Individuen zutreffen.
Regel 8 weist eine Bedingung für die rVO_{2peak} von höchstens 49.9 ml/kg/min sowie für die Lak_{peak} eine Höhe von maximal 6.25 mmol/l auf. Die Regel trifft auf 1 Individuum aus der Gruppe low-Boxen, 11 Individuen aus der Gruppe low-Fußball und auf 1 Individuum aus der Gruppe low-Handball zu. Somit ist die Regel auf die Wertausprägungen von 13 Individuen aus Cluster low anwendbar. Die Altersstruktur der betroffenen Individuen liegt mit 14, 15, 16 und 17 Jahre im jungen bis unteren mittleren Bereich. Ein Bezug der Regel zu Sportart oder Alter kann hier nicht mit Sicherheit hergestellt werden. Auffällig ist jedoch das eher jüngere Alter der Individuen. In Bezug auf die tlim kann folgende Aussage für die 8. Regel getroffen werden. Mit einer rVO_{2peak} unterhalb beziehungsweise gleich 49.9 ml/kg/min und einer Lak_{peak}-Ausprägung unterhalb beziehungsweise gleich 6.25 mmol/l ist keine tlim von 25 min zu erbringen.
Die Bedingungen von Regel 9 liegen in einer rVO_{2peak} von maximal 49.9 ml/kg/min. Darüber hinaus liegen bei der Regel die Werte für die Lak_{peak} bei über 6.75 mmol/l. Die Werte des Parameters Hb liegen zwischen einschließlich 12.2 und maximal 13.55 g/dl. Die Regel trifft auf die Werte von 5 Individuen aus der Gruppe low-Fußball zu. Die Altersstruktur liegt bei 15, 16 und 17 Jahren. In Zusammenhang mit der tlim kann die folgende Aussage für Regel 9 getroffen werden. Für Laufzeiten zwischen 18 und 20 min reicht eine rVO_{2peak} von maximal 49.9 ml/kg/min aus, wenn der Wert für die Lak_{peak} über 6.75 mmol/l liegt und der Wert für den Hb mindestens 12.2 und maxmimal 13.55 g/dl beträgt.
Regel 10 umfasst Bedingungen für die Parameter rVO_{2peak}, RQ_{peak}, Lak_{peak} und Hb. Dabei ist die rVO_{2peak} größer als 49.9 ml/kg/min, der RQ_{peak} über 1.075, die Lak_{peak} höchstens 7.1 mmol/l und der Hb höchstens 13.55 g/dl. Die Regel ist für die Werte von 3 Individuen aus der Gruppe low-Fußball mit einer Altersstruktur von 14 und 15 Jahren anwendbar. In Bezug auf die

Laufzeit kann die folgende Aussage getätigt werden. Für die Erbringung einer Laufzeit tlim von maximal 22 min kann die rVO_{2peak} über 49.9 ml/kg/min liegen, wenn Lak_{peak} höchstens 7.1 mmol/l beträgt und Hb einen Wert von 13.55 g/dl nicht überschreitet.

Werden die Ausprägungen der Lak_{peak} und der tlim in Bezug zueinander gesetzt, können innerhalb der verschiedenen Regeln Verdichtungen dieser Ausprägungskombinationen gefunden werden. Abb. 35 auf S. 127 zeigt ein Streudiagramm mit den Ausprägungskombinationen der Individuen für die Parameter Lak_{peak} und tlim. Auf der x-Achse des Diagramms sind die Werte für

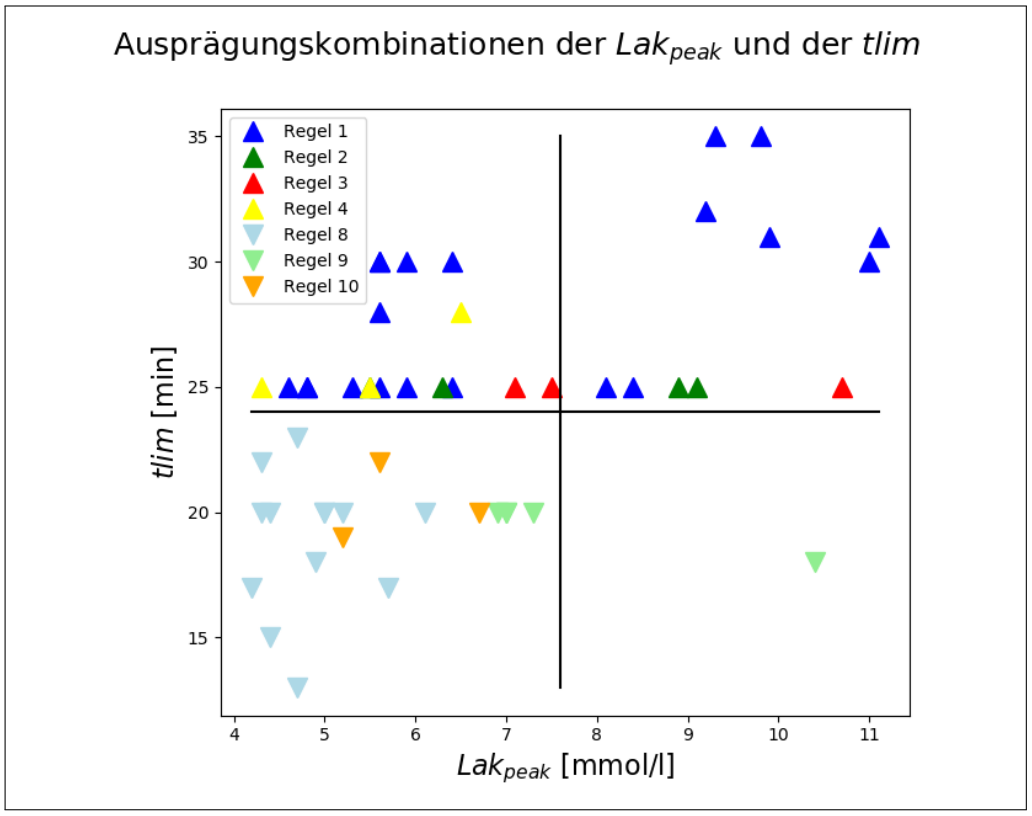

Abbildung 35: Ausprägungskombinationen der Lak_{peak} und der tlim nach Regeln (Quelle: Eigene Darstellung)

die Lak_{peak} auf einer einstufigen Skala von 4 bis 11 mmol/l verzeichnet. Die Werte für die tlim sind auf der y-Achse verzeichnet. Die dazugehörige fünf-

stufige Skala reicht von 15 bis 35 min. Des Weiteren ist das Diagramm durch Linien in Quadranten unterteilt. Die horizontale Linie ist auf einer Höhe von 24 min bei der tlim und auf einer Strecke von 4.2 bis 11.1 mmol/l bei der Lak_{peak} gezeichnet und teilt die Individuen in Cluster high und low. Durch die vertikale Linie wird die Lak_{peak}-Anhäufung während des Stufentests in zwei gleich große Hälften geteilt. Die Linie verläuft bei 7.6 mmol/l bei der Lak_{peak} und von 13 bis 35 min bei der tlim. Individuen aus Cluster high, für welche die Regeln 1, 2, 3 und 4 gelten, sind durch dunkelblaue, dunkelgrüne, rote und gelbe Dreiecke mit nach oben zeigender Spitze abgebildet. Individuen aus Cluster low, auf die die Regeln 8, 9 und 10 zutreffen, sind durch hellblaue, hellgrüne sowie orangenfarbige Dreiecke mit nach unten zeigender Spitze repräsentiert. Darüber hinaus kann dem Diagramm entnommen werden, welche Regeln für die einzelnen Individuen bezüglich der Parameter Lak_{peak} und tlim gelten. Bei näherer Betrachtung wird augenscheinlich, dass ein Großteil der hier abgebildeten Individuen aus Cluster low keine Lak_{peak}-Werte über 7.1 mmol/l aufweisen. Diese Aussage entspricht den Regeln 8 und 10, welche für 16 Individuen gelten. Regel 9 ist hingegen durchlässig für Lak_{peak}-Werte über 6.75 mmol/l. Die hellgrünen nach unten gerichteten Dreiecke kennzeichnen die entsprechenden Wertekombinationen. Bis auf eine Ausnahme befinden sich alle hier dargestellten Individuen aus Cluster low im unteren linken Quadranten. Bei diesen Individuen ist aufgrund der Ausprägung des Lak_{peak}-Wertes anzunehmen, dass der Zugriff auf die anaerobe Energiebereitstellung nicht so ausgeprägt ist, wie dies bei einem Teil der im Diagramm abgebildeten Individuen aus Cluster high der Fall ist. Bei diesen Individuen aus Cluster high kann davon ausgegangen werden, dass die anaerobe Energiebereitstellung ausgeprägter ist. Diese Annahme wird durch den Umstand gestützt, dass sich nahezu ein Drittel bis die Hälfte aller dieser Individuen aufgrund ihrer Ausprägungen in den Parametern Lak_{peak} und tlim im oberen rechten Quadranten befinden. Für Cluster high ist zu beobachten, dass die Regeln heterogener sind. Es gibt keine so klaren Grenzen bezüglich der Wertekombinationen von Lak_{peak} und tlim wie bei den Regeln für Cluster low. Dies kann an den Regeln 1, 2 und 3 beobachtet werden. Die Individuen, auf die diese Regeln zutreffen, sind nicht regelweise auf einen

Quadranten aufgeteilt. Eine Ausnahme bildet hier Regel 4. Alle Individuen, auf die diese Regel zutrifft, befinden sich im oberen linken Quadranten.

Darüber hinaus können anhand eines Streudiagramms wie dem in Abb. 35 auf S. 127 Individuen mit extremen Ausprägungen an Parameter-Kombinationen von Lak_{peak} und tlim identifiziert werden. 6 solcher Individuen sind innerhalb dieses Streudiagramms auszumachen und in Abb. 36 auf S. 129 farblich hervorgehoben. Auf der x-Achse des Streudiagramms sind die Werte

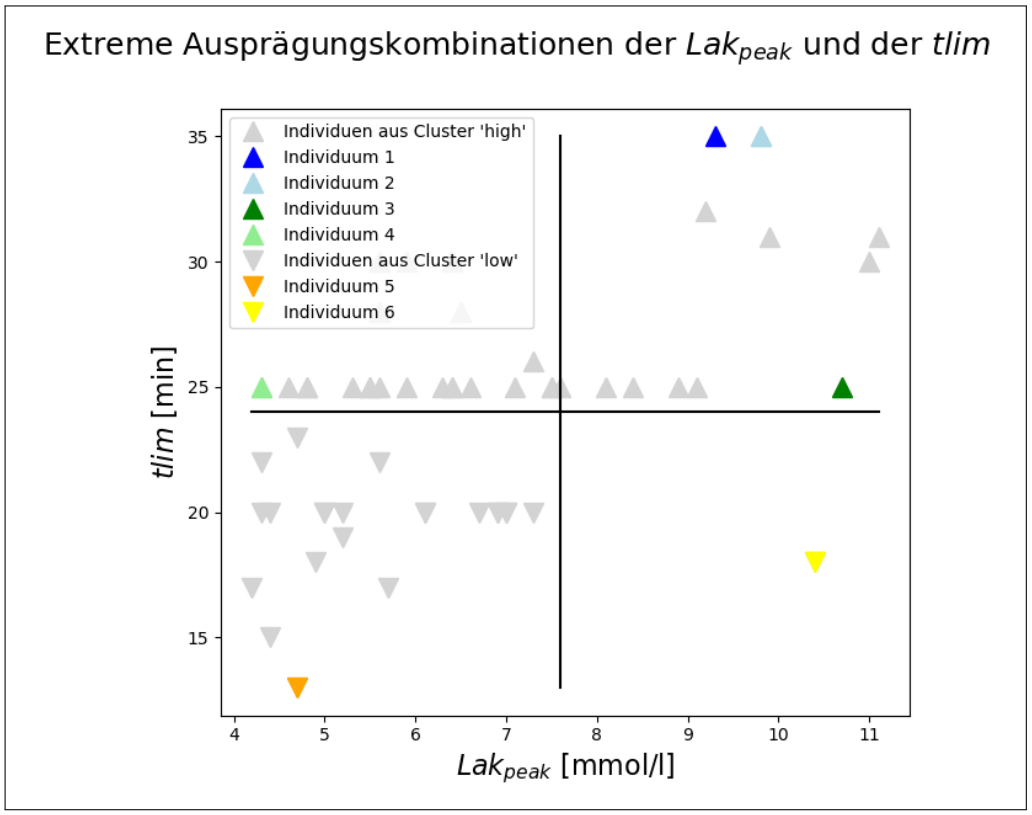

Abbildung 36: Extreme Ausprägungskombinationen der Lak_{peak} und der tlim nach Regeln (Quelle: Eigene Darstellung)

für die Lak_{peak} auf einer einstufigen Skala von 4 bis 11 mmol/l verzeichnet. Die Werte für die tlim sind auf der y-Achse zu finden. Die dazugehörige fünfstufige Skala reicht von 15 bis 35 min. Des Weiteren ist auch dieses Diagramm durch Linien in Quadranten unterteilt. Die horizontale Linie ist auf einer Höhe von 24 min bei der tlim und auf einer Strecke von 4.2 bis

11.1 mmol/l bei der Lak_{peak} gezeichnet und zeigt die Unterteilung der Individuen in Cluster high und low auf. Durch die vertikale Linie wird die Lak_{peak}-Anhäufung während des Stufentests in zwei gleich große Hälften geteilt. Die Linie verläuft bei 7.6 mmol/l bei der Lak_{peak} und von 13 bis 35 min bei der tlim. Die Wertekombinationen von der Lak_{peak} und der tlim für die einzelnen Individuen werden durch Dreiecke repräsentiert. Individuen aus Cluster high werden durch hellgraue Dreiecke dargestellt, deren Spitze nach oben zeigt. Individuen aus Cluster low werden durch hellgraue Dreiecke mit nach unten zeigender Spitze repräsentiert. Die erwähnten sechs Individuen sind durch farbige Dreiecke markiert. Die Individuen 1, 2, 3 und 4 stammen aus Cluster high und sind durch das dunkelblaue, das hellblaue, das dunkelgrüne sowie das hellgrüne Dreieck gekennzeichnet, deren Spitze jeweils nach oben zeigt. Individuum 5 und 6 aus Cluster low sind durch das orange beziehungsweise gelbe Dreieck mit nach unten deutender Spitze repräsentiert. Als extreme Wertausprägungen werden in diesem Zusammenhang solche Ausprägungen verstanden, durch die Individuen an den äußeren Randbereichen des Diagramms vertreten sind. Es können somit Individuen aus allen vier Quadranten identifiziert werden. Individuum 1, 2 und 3 befinden sich im oberen rechten Quadranten. Dabei weisen Individuum 1 und 2 die höchsten Werte für die tlim auf, jedoch nicht die höchsten im Parameter Lak_{peak}. Für Individum 3 kann hingegen ein sehr hoher Lak_{peak}-Wert verzeichnet werden, jedoch weist die tlim die niedrigste Ausprägung nicht nur im oberen rechten Quadranten, sondern auch für Cluster high insgesamt auf.
Entgegen Individuum 3 weist Individuum 4 aus dem oberen linken Quadranten einen der niedrigsten Lak_{peak}-Werte auf, obwohl es die gleiche Ausprägung für den Parameter tlim besitzt wie Individuum 3.
Das Individuum mit der geringsten Ausprägung für die tlim innerhalb des Diagramms ist Individuum 5. Es befindet sich aufgrund der niedrigen Ausprägungen von Lak_{peak} und tlim im unteren linken Quadranten. Eine ebenfalls relativ niedrige Ausprägung für den Parameter tlim weist Individuum 6 auf. Bei ihm ist jedoch eine recht hohe Ausprägung für die Lak_{peak} vorhanden. Somit befindet sich das Individuum aufgrund seiner Wertausprägungen in dem unteren rechten Quadranten.

Im weiteren Verlauf dieses Kapitels werden die Wertekombinationen dieser sechs Individuen einer genaueren Betrachtung unterzogen. Zu diesem Zweck sind die Wertausprägungen der verschiedenen Parameter für die einzelnen Individuen sowohl als Parallelkoordinatendiagramm in Abb. 37 auf S. 131 als auch in Tabelle 14 auf S. 131, Tabelle 15 auf S. 132 und Tabelle 16 auf S. 132 dargestellt.

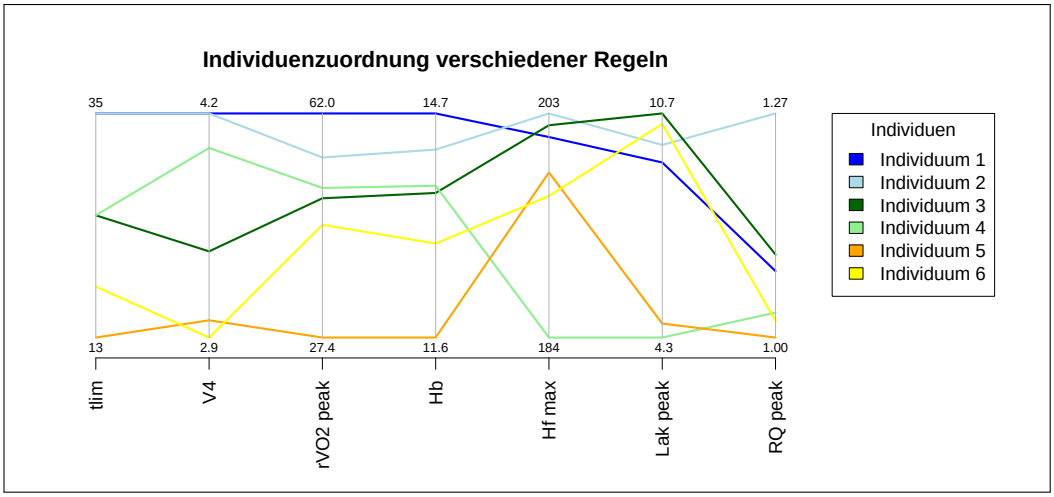

Abbildung 37: Extreme Parameterausprägungen einzelner Individuen (Quelle: Eigene Darstellung)

Individuum	Gruppe	Regel	Alter [Jahre]	tlim [min]	V4 [m/s]
1	high-Boxen	1	20	35	4.2
2	high-Boxen	1	16	35	4.2
3	high-Fußball	3	15	25	3.4
4	high-Fußball	4	15	25	4
5	low-Fußball	8	16	13	3
6	low-Fußball	9	17	18	2.9

Tabelle 14: Individuenwerte für Alter, tlim und V4

Aufgrund der Wertausprägungen für die tlim und teilweise auch die V4 können die Individuen wiederum in drei Individuengruppen unterteilt werden. Die Gruppe mit der höchsten Laufzeit besteht aus den Individuen 1 und 2 und basiert auf den höchsten Werten der tlim von 35 min. Individuum

Individuum	Gruppe	rVO_{2peak} [ml/kg/min]	Hb [g/dl]	Hf_{max} [S/min]
1	high-Boxen	62	14.7	201
2	high-Boxen	55.2	14.2	203
3	high-Fußball	48.9	13.6	202
4	high-Fußball	50.5	13.7	184
5	low-Fußball	27.4	11.6	198
6	low-Fußball	44.8	12.9	196

Tabelle 15: Individuenwerte für rVO_{2peak}, Hb und Hf_{max}

Individuum	Gruppe	Lak_{peak} [mmol/l]	RQ_{peak}
1	high-Boxen	9.3	1.08
2	high-Boxen	9.8	1.27
3	high-Fußball	10.7	1.1
4	high-Fußball	4.3	1.03
5	low-Fußball	4.7	1
6	low-Fußball	10.4	1.02

Tabelle 16: Individuenwerte für Lak_{peak} und RQ_{peak}

3 und 4 können jeweils aufgrund eines Wertes für die tlim von 25 min als Gruppe mit einer mittleren tlim zusammengefasst werden. Die Gruppe mit den niedrigsten Laufzeiten mit einer tlim von 13 beziehungsweise 18 min besteht aus den Individuen 5 und 6. Es ist zu beobachten, dass die Laufzeiten – auch innerhalb der gleichen Individuengruppe – auf voneinander abweichenden Ausprägungen innerhalb der physiologischen Parameter basieren.
So weist Individuum 1 die höchsten Werte für die rVO_{2peak} und den Hb auf. Individuum 2 besitzt für die tlim zwar die gleiche Ausprägung von 35 min, jedoch einen um 6.8 ml/kg/min niedrigeren rVO_{2peak}-Wert. Ebenso kann ein um 0.5 g/dl niedrigerer Wert für Hb festgestellt werden. Die Differenz der Hf_{max} liegt bei nur 2 S/min und ist für Individuum 1 mit 201 S/min niedriger. Beim Vergleich der Lak_{peak}-Werte ist festzustellen, dass Individuum 1 einen um 0.5 mmol/l niedrigeren Wert für die Lak_{peak} aufweist. Somit ist eine höhere anaerobe Energiebereitstellung bei Individuum 2 für die Erbringung der Laufzeit erfolgt als bei Individuum 1. Dies kann auch

anhand des RQ_{peak} festgestellt werden. Dieser zeigt deutlich, dass Individuum 2 stärker Kohlenhydrate zur Energiegewinnung verstoffwechselt hat. Da hier ein RQ_{peak} von über 1.2 vorliegt, kann sogar davon ausgegangen werden, dass von Individuum 2 eine anaerobe Schuld[101] eingegangen wurde. Dieses Individuum scheint eine hohe Motivation während des Laufbandtests gehabt zu haben. Es können somit zwei unterschiedliche Ausprägungen bei der Beanspruchung des Metabolismus zur Energiebereitstellung beobachtet werden, um die gleiche Laufzeit zu erbringen. Zum einen können mehr Fettsäuren mit Hilfe von O_2 zur Energiebereitstellung verstoffwechselt werden, zum anderen kann eine solche Laufzeit auch mit einer stärkeren Ausprägung des Energiestoffwechsels durch Glykose unter einer O_2-Schuld erbracht werden.
Die Individuen 3 und 4 weisen in dem Parameter tlim beide eine Ausprägung von 25 min auf. Die Werte von V4 liegen jedoch im Vergleich zu den jeweiligen Werten der Individuen 1 und 2 mit einer Differenz von 0.6 m/s relativ weit auseinander. Des Weiteren liegen die Ausprägungen für die $rVO_{2_{peak}}$ und den Hb mit 1.6 ml/min/kg beziehungsweise 0.1 g/dl relativ nah beieinander. Eine größere Differenz ist jedoch bei dem Parameter Hf_{max} auszumachen. Hier existiert eine Differenz von 18 S/min. Da die Werte beider Individuen für die $rVO_{2_{peak}}$ jedoch kaum differieren, könnte die große Differenz bei den Ausprägungen der Hf_{max} auf ein unterschiedlich stark ausgeprägtes Herzauswurfvolumen hindeuten. Auch beim Wert der Lak_{peak} existiert eine große Differenz von 6.4 mmol/l zwischen den beiden Individuen. Eine Differenz ist auch bei den jeweiligen Ausprägungen des Parameters RQ_{peak} zu erkennen. Diese Ausprägungen betragen 1.1 bei Individuum 3 und 1.03 bei Individuum 4. Somit ist zu erkennen, dass bei Individuum 4 die Energiebereitstellung mehr auf der Basis von Fettsäuren als auf der Basis von Glykose stattgefunden hat. Bei Individuum 3 scheint dies anders zu sein. Hier ist die Energiebereitstellung durch Glykose stärker ausgeprägt gewesen als durch Fettsäuren.
Die beiden Individuen 5 und 6 aus der Gruppe low-Fußball weisen mit 3 be-

[101] Siehe auch Kapitel 2.3.2 ab S. 35.

ziehungsweise 2.9 m/s jeweils einen sehr ähnlichen Wert für die V4 auf. Bei der Laufzeit unterscheiden sich die Werte um 5 min. Individuum 5 weist mit 13 min die niedrigste tlim auf, Individuum 6 mit 18 min die zweitniedrigste. Die unterschiedlichen Wertausprägungen für die tlim können durch die ebenfalls unterschiedlichen Ausprägungen des Parameters $rVO_{2_{peak}}$ begründet werden. Dieser liegt für Individuum 5 bei 27.4 ml/kg/min. Der niedrige Wert für die $rVO_{2_{peak}}$ kann wiederum durch den niedrigen Hb-Wert von 11.6 g/dl erklärt werden. Der im Gegensatz dazu höhere Wert von Individuum 6 für den Hb von 12.9 g/dl macht wiederum den höheren Wert von 44.8 ml/kg/-min für die $rVO_{2_{peak}}$ plausibel. Da die Hf_{max} bei beiden Individuen mit 198 beziehungsweise 196 S/min ungefähr gleich hoch ausfällt, kann ein Einfluss der Hf_{max} auf die unterschiedlichen Ausprägungen der tlim hier als unwahrscheinlich angenommen werden. Die Lak_{peak}-Werte von Individuum 5 und 6 unterscheiden sich um 5.7 mmol/l bei einem nahezu identischen RQ_{peak}-Wert von 1 beziehungsweise 1.02.

Zusammenfassend kann festgestellt werden, dass ein direkter Zusammenhang zwischen den Parametern tlim sowie $rVO_{2_{peak}}$ und Hb besteht. Darüber hinaus deuten die Regeln auf unterschiedliche Ausprägungen von aerober und anaerober Energiebereitstellung bei den hier betrachteten Individuen hin.

6 Fazit und Ausblick

Das letzte Kapitel dieser Arbeit enthält das Fazit als Zusammenfassung und Bewertung der Ergebnisse dieser Arbeit. Im daran anschließenden Ausblick werden die Potentiale des im Rahmen dieser Arbeit entstandenen Systems sowie des analytischen Modells aufgeführt.

6.1 Fazit

Innerhalb der Sportwissenschaft und speziell im Bereich der Ausdauerdiagnostik und des Ausdauertrainings besteht aktuell die Notwendigkeit der Entwicklung und Anwendung analytischer Modelle, die zur Entscheidungsunterstützung für ein gezieltes individuelles Training eingesetzt werden können. Solche Modelle zur Entscheidungsunterstützung beruhen auf Daten, die teilweise an ganz anderen Stellen einer Organisation erhoben werden als die aus diesen entstehenden analytischen Modelle. Insofern ist ein übergeordneter Prozess wie der von Ward et al. (2019) geforderte notwendig, um die Integration, Persistierung und Auswertung dieser Daten sowie die Kommunikation der aus den Daten gewonnenen Kenntnisse zu gewährleisten.

Im Rahmen der vorliegenden Arbeit wurde ein solcher Prozess durch den Data-Warehouse-Prozess[102] und den CRISP-DM[103] konkretisiert. Dabei wurde der Unterprozess der Datenbeschaffung mit Hilfe einer innerhalb dieser Arbeit als Prototyp nach spezifizierten Anforderungen[104] implementierten Server-/Client-Applikation[105] realisiert, die die Datenpersistierung in eine relationale Datenbank vollzog. Der Unterprozess der Datenanalyse wurde durch den CRISP-DM abgebildet. Durch dessen iterative Vorgehensweise konnte aus spiroergometrischen sowie hämatologischen Parametern von Datensätzen[106] verschiedener Individuen unterschiedlichen Alters, Geschlechts und

[102] Siehe auch Kapitel 2.1.1 ab S. 14.
[103] Siehe auch Kapitel 2.1.2 ab S. 19.
[104] Siehe Kapitel 4.1 ab S. 51.
[105] Siehe Kapitel 4.3 ab S. 57.
[106] Siehe Kapitel 3.2 ab S. 47.

ausgeübter Sportarten ein analytisches Modell auf der Basis eines agglomerativen Clusterings[107] sowie eines auf diesem aufbauenden Decision Trees[108] erstellt werden. Dabei diente das Clustering dazu, die Individuen in Leistungscluster basierend auf den Parametern tlim und V4 einzuteilen.[109] Die durch das Clustering gefundenen Leistungscluster wurden danach als Klassen für den Decision Tree verwendet, dem daraufhin Regeln für die physiologischen Parameter $rVO_{2_{peak}}$, RQ_{peak}, Hf_{max}, Lak_{peak} und Hb entnommen werden konnten.[110]

Die beiden innerhalb dieser Arbeit abgebildeten Prozesse stellten den notwendigen Rahmen für das Vorgehen der Datenintegration und -analyse bereit, ermöglichten ein strukturiertes Vorgehen und bildeten darüber hinaus die Grundlage für das implementierte Client-/Server-System.
Das System selbst ist sehr schlank und stellt auf der Seite des Servers eine JSON-API bereit, über die sowohl Datenbanktabellen in einer relationalen Datenbank angelegt als auch Datensätze in diese integriert werden können. Die dem System zugrunde liegende relationale Datenbank ermöglichte sowohl eine Verknüpfung der spiroergometrischen mit den hämatologischen Daten als auch die Auswahl der Datensätze, auf die die einzelnen Regeln des erstellten analytischen Modells zutrafen. Der Client des entwickelten Systems ermöglicht jedoch keine automatisierte Datenaggregation, weshalb eine solche manuell durchgeführt werden musste.
Das analytische Modell läßt aufgrund der parameterübergreifenden Regeln Rückschlüsse auf Ausprägungen von Parameterkombinationen zu. Damit ermöglichen die in dieser Arbeit erstellten Regeln einen exakteren Einblick in die physiologischen Ausprägungen der Leistungscluster als dies beispielsweise durch Streudiagramme zwischen nur zwei Parametern möglich wäre. Sportwissenschaftler und Betreuungspersonal erhalten mit Hilfe des Modells konkrete Wertekombinationen für Parameter, mit denen eine bestimmte tlim

[107] Siehe auch Kapitel 2.2.1 ab S. 22.
[108] Siehe auch Kapitel 2.2.2 ab S. 27.
[109] Siehe auch Kapitel 2.3.2 ab S. 35.
[110] Siehe auch Kapitel 2.3.2 ab S. 35.

erreicht werden kann. Dabei ist das Modell unabhängig auf das Geschlecht, Alter oder die Sportart der jeweiligen Individuen anwendbar. Des Weiteren können anhand der Regeln Auffälligkeiten entdeckt werden, wie beispielsweise die Einordnung eines Individuums in ein Cluster, das aufgrund der Parameterausprägungen zunächst als unwahrscheinlich gelten würde. Das Modell wurde auf der Basis zweier Skripte erstellt, die beginnend mit dem Clustering sequenziell ausgeführt wurden. Der Informationsaustausch zwischen den beiden Skripten erfolgte durch den manuellen Transfer einer CSV-Datei, was zu einem Medienbruch und erhöhtem Aufwand bei der Datenanalyse führte. Durch das Clustering wurden die Individuen in zwei Leistungscluster mit einer Grenze bei der tlim von 24 min unterteilt. Die Ausprägungen der jeweiligen physiologischen Parameter aus den beiden Leistungsclustern wurden mit Hilfe des Decision Trees beziehungsweise der aus diesem resultierenden Regeln abgeleitet. Aufgrund der Art und Weise, wie der Decision-Tree-Algorithmus die Regeln berechnet, sind die Grenzen für die einzelnen Parameterausprägungen, die nicht den in den Datensätzen vorkommenden Ausprägungen entsprechen, berechneter Natur.
Des Weiteren sind die Grenzen, die durch die verwendeten Algorithmen sowohl für die Leistungscluster als auch für die Regeln berechnet wurden, keine Intervalle. Solche Intervalle als fließende Grenzen entsprächen eventuell eher den physiologischen Gegebenheiten.
Innerhalb des erstellten Modells existieren Regeln, die sich bezogen auf die Erbringung der tlim teilweise widersprechen. Ein solcher Widerspruch könnte eventuell unterschiedliche Ausprägungen physiologischer Natur aufzeigen.

Zusammenfassend kann hier angemerkt werden, dass die Ziele dieser Arbeit erreicht werden konnten. Der Data-Warehouse-Prozess und der CRISP-DM können erfolgreich in der Sportwissenschaft angewendet werden. Dabei kann ein zweistufiges auf ML basierendes Modell, das eine Einteilung in Leistungscluster und eine Analyse der phyisologischen Strukturen dieser Cluster ermöglicht, mit Erfolg eingesetzt werden.
ML kann somit erfolgreich zur Analyse in der Sportwissenschaft angewendet werden, wie auch Yeung (2018) aufführt. Einen Trainer zu ersetzen, wie dies

beispielsweise bei Köpf (2019) angeführt wird, erscheint mit Hilfe des in dieser Arbeit erstellten Modells nicht möglich, da die Ergebnisse des Modells im jeweiligen Kontext interpretiert werden müssen. Das erstellte Modell ist jedoch innerhalb des in dieser Arbeit aufgeführten Rahmens in der Lage, einen Trainer bei der Entscheidungsfindung im Bereich der Ausdauerdiagnostik und des Ausdauertrainings zu unterstützen.

6.2 Ausblick

Sowohl bei der Betrachtung des innerhalb dieser Arbeit implementierten systemischen Prototyps als auch bei dem enstandenen analytischen Modell sind diverse Erweiterungen beziehungsweise die Verwendung anderer Algorithmen denkbar. Darüber hinaus können auch weitere Anwendungsfälle für das System und das analytische Modell in Betracht gezogen werden.

Um die Integrität der Datenstrukturen bei der Integration zu erhöhen, wäre eine JSON-Validierung über ein Schema möglich. Eine solche Valdierung könnte beispielsweise innerhalb eines Request-Controllers vor der weiteren Verarbeitung des im Request übertragenen JSON sein. Des Weiteren könnte das JSON-API nach der JSON:API Spezifikation (*Latest Specification (v1.0)*, o. J.) implementiert und somit standardisiert werden.
Weitere Optimierungen wären durch die Übertragung von HTTP-Status-Codes innerhalb der JSON-Response denkbar.
Auch wären Authentifizierungsmechanismen wie die Verwendung von Zertifikaten für produktive Einsätze unabdingbar.
Die Limitierungen wie beispielsweise die fehlende Aggregationsmöglichkeit des CSV-Clients bei der Datenintegration könnten durch die Verwendung entsprechender Datenintegrationswerkzeuge[111] vermieden werden. Welche Werkzeuge dafür potentiell geeignet sind, wäre in einem weiteren Schritt zu evaluieren.
Die im Prototypen vorhandenen Medienbrüche bei der Datenanalyse könnten ebenfalls in zukünftigen Entwicklungsschritten entfernt werden.
Um Ergebnisse aus dem analytischen Modell an System-Benutzer kommunizieren zu können, wäre eine Art Dashboard denkbar, wie in der Abbildung 38 auf S. 140 schematisch zu sehen ist. Bei einem solchen Dashboard könnten die durch das analytische Modell gefundenen Leistungscluster mit der Anzahl an Individuen (n) die höchste Ebene bilden. Unterhalb der einzelnen Cluster wären die jeweiligen Regeln gemäß ihrer Individuenanzahl (n) aufgeführt. Auf der Ebene unterhalb der Regeln wären dann die Individuen, auf welche

[111] Siehe auch Kapitel 1.2 ab S. 5.

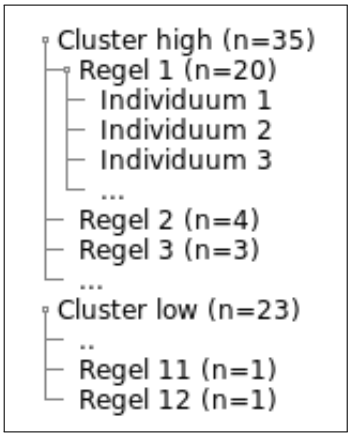

Abbildung 38: Wireframe eines Dashboards (Quelle: Eigene Darstellung)

die Regeln zutreffen, mit ihren konkreten Parameterausprägungen aufzulisten.

In Bezug auf das analytische Modell könnten weitere Algorithmen auf ihre Eignung untersucht werden. Interessant wäre hier die Evaluation von Neuronalen Netzen, SVMs, Random Forest sowie Association Rules.
Des Weiteren wäre auch die Untersuchung von Datensätzen einzelner ausgewählter Individuen mit einer besonders hohen tlim mittels der hier erwähnten Verfahren in Erwägung zu ziehen, um weitere Einblicke in die physiologischen Strukturen zu erhalten, welche Strukturen Höchstleistungen im Ausdauerbereich ermöglichen.
Neben diesen weiterführenden Ansätzen wäre zukünftig auch ein Prozess denkbar, der die für die Analyse notwendigen Daten mit Hilfe von Wearables erfasst. Ein solches Wearable könnte beispielsweise ein Muscle Oxiometer sein, das die O_2-Zufuhr und Hämodynamik innerhalb von Muskelgeweben misst (Lutz, Memmert, Raabe, Dornberger & Donath, 2019).

Zukünftige Anwendungen des in diesem Unterkapitel beschriebenen Systems sowie des analytischen Modells könnten Trainingspersonal bei der Entscheidungsfindung in der Trainingsplanung unterstützen, indem speziell auf einzelne Cluster und die für diese geltenden Regeln bezogene Trainingspläne erstellt

werden könnten. Konkret könnten die hier beschriebenen Vorgehensweisen und Verfahren so in Sportkadern wie beispielsweise einem Fußballbundesliga-Verein eingesetzt werden.

Literatur

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems* (Bericht). Zugriff am 06.11.2020 auf `https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf`

About the Initiative. (o. J.). DigiTwins. Zugriff am 06.11.2020 auf `https://www.digitwins.org/about-the-initiative`

Abut, F. & Akay, M. F. (2015). Machine learning and statistical methods for the prediction of maximal oxygen uptake: recent advances. *Medical Devices: Evidence and Research*, *2015* (8), 369-379. doi: 10.2147/mder.s57281

Abut, F., Akay, M. F. & George, J. (2016). Developing new VO_2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in Biology and Medicine*, *79*, 182 - 192. doi: 10.1016/j.compbiomed.2016.10.018

Ahrens, P. (2018). *Technikeinsatz bei der WM 2018 – Generation Tablet-Trainer.* Der Spiegel. Zugriff am 06.11.2020 auf `https://www.spiegel.de/sport/fussball/fussball-wm-2018-joachim-loew-warum-fuer-trainer-technik-wichtiger-wird-a-1210390.html`

AI and machine learning products. (o. J.). Google Cloud. Zugriff am 06.11.2020 auf `https://cloud.google.com/products/ai`

Anaconda (Version 3.5.1) [Computerprogramm]. (2018). Anaconda Software Distribution. Zugriff am 06.11.2020 auf `https://repo.anaconda.com/archive/Anaconda3-5.1.0-Linux-x86_64.sh`

Apache Commons CSV (Version 1.5) [Computerprogramm]. (o. J.). The Apache Software Foundation. Zugriff am 14.12.2020 auf `https://archive.apache.org/dist/commons/csv/binaries/commons-csv-1.5-bin.zip`

Apache Commons Lang (Version 3.4) [Computerprogramm]. (o. J.). The Apache Software Foundation. Zugriff am 14.12.2020 auf

https://archive.apache.org/dist/commons/lang/binaries/commons-lang3-3.4-bin.zip

Apache Kafka® is a distributed streaming platform. What exactly does that mean? (o. J.). Apache Kafka. Zugriff am 06.11.2020 auf https://kafka.apache.org/intro#intro_platform

Apache Maven (Version 3.6.0) [Computerprogramm]. (2018). Apache Software Foundation Distribution Directory. Zugriff am 07.11.2020 auf https://archive.apache.org/dist/maven/maven-3/3.6.0/binaries/apache-maven-3.6.0-bin.zip

Apache NiFi Overview. (2020). Apache nifi. Zugriff am 06.11.2020 auf https://nifi.apache.org/docs.html

Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse* (3., ergänzte überarbeitete und neu gestaltete Aufl.). München: Oldenbourg Wissenschaftsverlag.

Bain, B. J., Bates, I., Laffan, M. A. & Lewis, S. M. (2012). *Dacie and Lewis Practical Haematology* (11. Aufl.). Edinburgh: Elsevier Churchill Livingstone.

Bauer, A. & Günzel, H. (Hrsg.). (2013). *Data Warehouse Systeme* (4., überarbeitete und erweiterte Aufl.). Heidelberg: dpunkt.verl.

Beckert, S., Farrahi, F., Aslam, R. S., Scheuenstuhl, H., Königsrainer, A., Hussain, M. Z. & Hunt, T. K. (2006). Lactate stimulates endothelial cell migration. *Wound repair and regeneration : official publication of the Wound Healing Society [and] the European Tissue Repair Society*, *14* (3), 321-324. doi: 10.1111/j.1743-6109.2006.00127.x

Behme, W. & Mucksch, H. (Hrsg.). (2001). *Data Warehouse-gestützte Anwendungen* (1. Aufl.). Wiesbaden: Gabler Verlag.

Behrends, J. C., Bischofberger, J., Deutzmann, R., Ehmke, H., Frings, S., Grissmer, S., ... Wischmeyer, E. (2012). *Physiologie* (2., überarbeitete Aufl.). Stuttgart: Thieme Verlag.

Bellazzi, R. & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, *77* (2), 81–97. doi: 10.1016/j.ijmedinf.2006.11.006

Bellinger, I. & Brennan, C. (2018, Juli). Höher schneller schlauer. *National*

Geographic, 40-65.

Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., ... Swedish Digital Twin Consortium (2019). Digital twins to personalize medicine. *Genome Medicine*, *12* (1), 4. doi: 10.1186/s13073-019-0701-3

Boeselager, F. (2018). *Künstliche Intelligenz in der Bundesliga Talentscouting 4.0.* Deutschlandfunk. Zugriff am 06.11.2020 auf `https://www.deutschlandfunk.de/kuenstliche-intelligenz-in-der-bundesliga-talentscouting-4-0.1346.de.html?dram:article_id=428803`

Bonen, A. (2000). Lactate transporters (MCT proteins) in heart and skeletal muscles. *Medicine and Science in Sports and Exercise*, *32* (4), 778-789. doi: 10.1097/00005768-200004000-00010

Brachet, P. (2020). Texmaker (Version 5.0.4) [Software]. Zugriff am 06.11.2020 auf `https://www.xm1math.net/texmaker/assets/files/texmaker_5.0.4_ubuntu_18_04_amd64.deb`

Broich, H. (2009). *Quantitative Verfahren zur Leistungsdiagnostik im Leistungsfußball: Empirische Studien und Evaluationen verschiedener leistungsrelevanter Parameter* (Dissertation). Deutsche Sporthochschule Köln, Köln.

Brooks, G. A. (2007). Lactate. *Sports Medicine*, *37* (4-5), 341–343. doi: 10.2165/00007256-200737040-00017

Burtscher, M., Nachbauer, W. & Wilber, R. (2011). The upper limit of aerobic power in humans. *European Journal of Applied Physiology*, *111* (10), 2625–2628. doi: 10.1007/s00421-011-1885-4

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0* (Bericht). Zugriff am 06.11.2020 auf `ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf`

Chollet, F. et al. (2015). *Keras.* Zugriff am 06.11.2020 auf `https://keras.io`

CloverDX. (2020). SourceForge. Zugriff am 06.11.2020 auf `https://`

sourceforge.net/projects/cloveretl

Data Integration - Kettle. (2017). Hitachi. Zugriff am 06.11.2020 auf https://community.hitachivantara.com/s/article/data-integration-kettle

Deru, M. & Ndiaye, A. (2019). *Deep Learning mit TensorFlow, Keras und TensorFlow.js.* Bonn: Rheinwerk Verlag.

Dworschak, M. (2018). *DeepMind und Co. Was künstliche Intelligenz schon leisten kann - und was nicht.* Der Spiegel. Zugriff am 06.11.2020 auf https://www.spiegel.de/spiegel/was-kuenstliche-intelligenz-schon-leisten-kann-und-was-nicht-a-1186438.html

Eckert, S., Hurtz, S., Müller-Hansen, S. & Wormer, V. (2019). *Die Like-Fabrik.* Süddeutsche Zeitung. Zugriff am 06.11.2020 auf https://www.sueddeutsche.de/digital/paidlikes-gekaufte-likes-facebook-instagram-youtube-1.4728833!amp

Engelmeyer, E. (2012). *Analyse des Gesundheits- und Leistungsstatus von Kaderathleten olympischer Sportarten* (Dissertation). Deutsche Sporthochschule Köln, Hamburg.

Faeskorn-Woyke, H., Bertelsmeier, B., Riemer, P. & Bauer, E. (2007). *Datenbanksysteme.* München: Pearson Studium.

Faude, O., Kindermann, W. & Meyer, T. (2009). Lactate Threshold Concepts: How Valid are They? *Sports Medicine*, *39* (6), 469–490. doi: 10.2165/00007256-200939060-00003

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, *17* (3), 37–54. doi: 10.1609/aimag.v17i3.1230

Gladden, L. B. (2004). Lactate metabolism: a new paradigm for the third millennium. *Journal of Physiology*, *558* (Pt 1), 5-30. doi: 10.1113/jphysiol.2003.058701

Gliozzo, A., Ackerson, C., Bhattacharya, R., Goering, A., Jumba, A., Kim, S. Y., ... Ribas, M. (2017). *Building Cognitive Applications with IBM Watson Services: Volume 1 Getting Started.* IBM. Zugriff am 06.11.2020 auf http://www.redbooks.ibm.com/redbooks/pdfs/sg248387.pdf

Göbel, M. O. (2019). *KI im Sport: So findet künstliche Intelligenz die besten Fußballspieler.* Basecamp. Zugriff am 06.11.2020 auf `https://www.basecamp.digital/ki-im-sport-so-findet-kuenstliche-intelligenz-die-besten-fussballspieler/`

Groll, A., Ley, C., Schauberger, G. & Eetvelde, H. V. (2018). *Prediction of the FIFA World Cup 2018 – A randomforest approach with an emphasis on estimated teamability parameters.* Zugriff am 06.11.2020 auf `https://arxiv.org/pdf/1806.03208.pdf`

Helical Insight. (o. J.). GitHub. Zugriff am 06.11.2020 auf `https://github.com/helicalinsight/helicalinsight`

Hickson, R. C., Bomze, H. A. & Holloszy, J. O. (1977). Linear increase in aerobic power induced by a strenuous program of endurance exercise. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, *42* (3), 372-376. doi: 10.1152/jappl.1977.42.3.372

Hohmann, A., Lames, M., Letzelter, M. & Pfeiffer, M. (2020). *Einführung in die Trainingswissenschaft* (7., überarbeitete Aufl.). Wiebelsheim: Limbert.

Huang, S.-C., Casaburi, R., Liao, M.-F., Liu, K.-C., Chen, Y.-J., Fu, T.-C. & Su, H.-R. (2019). Noninvasive prediction of Blood Lactate through a machine learning-based approach. *Scientific Reports*, *9* (1), 2180. doi: 10.1038/s41598-019-38698-1

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing In Science & Engineering*, *9* (3), 90-95. doi: 10.1109/MCSE.2007.55

Inmon, W. H. (2002). *Building the Data Warehouse* (3. Aufl.). New York [u.a.]: Wiley.

IntelliJ IDEA Community (Version 2020.1) [Computerprogramm]. (2020). JetBrains. Zugriff am 06.11.2020 auf `https://download.jetbrains.com/idea/ideaIC-2020.1.tar.gz`

JabRef (Version 3.8.2) [Computerprogramm]. (2017). GitHub. Zugriff am 06.11.2020 auf `https://github.com/JabRef/jabref/releases/download/v3.8.2/JabRef-3.8.2.jar`

Jarke, M., Lenzerini, M., Vassiliou, Y. & Vassiliadis, P. (2003). *Fundamentals of Data Warehouses* (2., überarbeitete und erweiterte Aufl.).

Berlin [u.a.]: Springer.

Jaspersoft® ETL. (o. J.). Jaspersoft® Community. Zugriff am 06.11.2020 auf https://community.jaspersoft.com/project/jaspersoft-etl

Joyner, M. J. & Coyle, E. F. (2008). Endurance exercise performance: the physiology of champions. *Journal of Physiology*, *586* (1), 35-44. doi: 10.1113/jphysiol.2007.143834

Kaplan, A. & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62* (1), 15-25. doi: 10.1016/j.bushor.2018.08.004

Karelis, A. D., Marcil, M., Péronnet, F. & Gardiner, P. F. (2004). Effect of lactate infusion on M-wave characteristics and force in the rat plantaris muscle during repeated stimulation in situ. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, *96* (6), 2133-2138. doi: 10.1152/japplphysiol.00037.2004

KETL. (2015). SourceForge. Zugriff am 06.11.2020 auf https://sourceforge.net/projects/ketl/

Kimball, R. & Ross, M. (2013). *The Data Warehouse Toolkit* (3. Aufl.). Indianapolis, Indiana: Wiley.

King, T. (2019a). *Top 12 Free and Open Source ETL Tools for Data Integration.* Solutions Review. Zugriff am 06.11.2020 auf https://solutionsreview.com/data-integration/top-free-and-open-source-etl-tools-for-data-integration/

King, T. (2019b). *Top 18 free and Open Source Business Intelligence Tools.* Solutions Review. Zugriff am 06.11.2020 auf https://solutionsreview.com/business-intelligence/top-free-and-open-source-business-intelligence-software-tools/

KNIME Analytics Platform. (o. J.). KNIME. Zugriff am 06.11.2020 auf https://www.knime.com/knime-analytics-platform

Köpf, A. (2019). *KI als Fußball-Trainer: Maschinelles Lernen soll FC Liverpool verbessern.* GameStar. Zugriff am 06.11.2020 auf https://www.gamestar.de/artikel/ki-als-fussball-trainer,3352688

.html

Köppen, V., Saake, G. & Sattler, K.-U. (2012). *Data Warehouse Technologien*. Heidelberg [u.a.]: mitp.

Kroidl, R. F., Schwarz, S., Lehnigk, B. & Fritsch, J. (Hrsg.). (2015). *Kursbuch Spiroergometrie. Technik und Befundung leicht gemacht* (3., vollständig überarbeitete und erweiterte Aufl.). Stuttgart: Thieme.

Larose, D. T. (2005). *Discovering Knowledge in Data – An Introduction to Data Mining*. Hoboken: Wiley. Zugriff am 09.11.2011 auf http://www.ce.sharif.edu/~valipour/to_be_read/DMBook.pdf

Latest Specification (v1.0). (o. J.). JSON API. Zugriff am 06.11.2020 auf https://jsonapi.org/format/

LibreOffice (Version 6.0.7.3) [Computerprogramm]. (2018). LibreOffice. Zugriff am 06.11.2020 auf https://downloadarchive.documentfoundation.org/libreoffice/old/6.0.7.3/deb/x86_64/LibreOffice_6.0.7.3_Linux_x86-64_deb.tar.gz

Link, D. (2018). *Data Analytics in Professional Soccer*. Wiesbaden: Springer Vieweg.

Linke, D., Link, D. & Lames, M. (2018). Validation of electronic performance and tracking systems EPTS under field conditions. *PLoS One*, *13* (7), e0199519. doi: 10.1371/journal.pone.0199519

Linux Mint (Version 19.1) [Computerprogramm]. (2018). Zugriff am 06.11.2020 auf http://mirror.bauhuette.fh-aachen.de/linuxmint-cd/stable/19.1/linuxmint-19.1-cinnamon-64bit.iso

Liu, B. (2011). *Web Data Mining* (2. Aufl.). Heidelberg [u.a.]: Springer.

Lundby, C. & Robach, P. (2015). Performance Enhancement: What Are the Physiological Limits? *Physiology (Bethesda)*, *30* (4), 282-292. doi: 10.1152/physiol.00052.2014

Lusti, M. (2002). *Data Warehousing und Data Mining* (2., überarbeitete und erweiterte Aufl.). Berlin/Heidelberg: Springer.

Lutz, J., Memmert, D., Raabe, D., Dornberger, R. & Donath, L. (2019). Wearables for Integrative Performance and Tactic Analyses: Opportunities, Challenges, and Future Directions. *International Jour-*

nal of Environmental Research and Public Health, *17* (1), 59. doi: 10.3390/ijerph17010059

Machine Learning in AWS. (o. J.). AWS. Zugriff am 06.11.2020 auf `https://aws.amazon.com/de/machine-learning/`

Mader, A., Liesen, H. & Heck, H. (1976). Zur Beurteilung der sportartspezifischen Ausdauerleistungsfähigkeit im Labor. *Sportarzt und Sportmedizin*, *27* (4), 80-88.

Memmert, D., Lemmink, K. A. P. M. & Sampaio, J. (2016). Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sports Medicine*, *47* (1), 1-10. doi: 10.1007/s40279-016-0562-5

Metabase. (2020). GitHub. Zugriff am 06.11.2020 auf `https://github.com/metabase/metabase/blob/master/README.md`

Milovanova, T. N., Bhopale, V. M., Sorokina, E. M., Moore, J. S., Hunt, T. K., Hauer-Jensen, M., ... Thom, S. R. (2008). Lactate Stimulates Vasculogenic Stem Cells via the Thioredoxin System and Engages an Autocrine Activation Loop Involving Hypoxia-Inducible Factor 1. *Molecular and Cellular Biology*, *28* (20), 6248-6261. doi: 10.1128/MCB.00795-08

Moeser, J. (2019). *Künstliche Intelligenz in der Bundesliga – SV Werder Bremen nutzt die intelligente Scouting Plattform SCOUTASTIC.* JAAI. Zugriff am 06.11.2020 auf `https://jaai.de/kuenstliche-intelligenz-in-der-bundesliga-sv-werder-bremen-nutzt-die-intelligente-scouting-plattform-jaai-scout-2082`

Müller, J. (2000). *Transformation operativer Daten zur Nutzung im Data Warehouse* (Dissertation). Universität Bochum, Wiesbaden.

Müllner, D. (2011). *Modern hierarchical, agglomerative clustering algorithms.* Zugriff am 06.11.2020 auf `https://arxiv.org/pdf/1109.2378.pdf`

MySQL Community Server (Version 5.7.29) [Computerprogramm]. (o. J.). MySQL. Zugriff am 06.11.2020 auf `https://dev.mysql.com/get/Downloads/MySQL-5.7/mysql-server_5.7.29-1ubuntu18.04_amd64.deb-bundle.tar`

MySQL Connector/J (Version 8.0.8-dmr) [Computerprogramm]. (o. J.).

MySQL. Zugriff am 06.11.2020 auf https://downloads.mysql.com/archives/get/p/3/file/mysql-connector-java_8.0.8-dmr-1ubuntu17.04_all.deb

Nöll, N. (2009). *Datenbankgestützte Informationstechnologie zur Unterstützung multidisziplinärer Forschung in der Sportwissenschaft* (Dissertation). Deutsche Sporthochschule Köln, Köln.

OpenJDK (Version 11.0.7) [Computerprogramm]. (o. J.). GitHub. Zugriff am 06.11.2020 auf https://github.com/AdoptOpenJDK/openjdk11-binaries/releases/download/jdk-11.0.7%2B10/OpenJDK11U-jdk_x64_linux_hotspot_11.0.7_10.tar.gz

Open Studio for Data Integration. (o. J.). Talend. Zugriff am 06.11.2020 auf https://www.talend.com/products/data-integration/data-integration-open-studio/

Papula, L. (2001). *Mathematik für Ingenieure und Naturwissenschaftler Band 3* (4., verbesserte Aufl.). Braunschweig/Wiesbaden: Vieweg.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

PlantUML (Version 1.2019.11) [Computerprogramm]. (2019). SourceForge. Zugriff am 06.11.2020 auf https://sourceforge.net/projects/plantuml/files/plantuml.1.2019.11.jar/download

Postman (Version 7.22.1) [Computerprogramm]. (o. J.). Postman. Zugriff am 12.04.2020 auf https://dl.pstmn.io/download/latest/linux64

PyCharm Community Edition (Version 2018.1.1) [Computerprogramm]. (o. J.). JetBrain. Zugriff am 06.11.2020 auf https://download.jetbrains.com/python/pycharm-community-2018.1.1.tar.gz

Raschka, S. (2017). *Maschine Learning mit Pyhton.* Frechen: mitp.

Rauch, G. & Litzel, N. (2016). *Was ist Watson?* BigData-Insider. Zugriff am 06.11.2020 auf https://www.bigdata-insider.de/was-ist-watson-a-572251/

Rein, R. & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Sprin-*

gerPlus, *5* (1), 1410. doi: 10.1186/s40064-016-3108-2

Revolutionise Sport Through AI. (o. J.). Stats Perform. Zugriff am 06.11.2020 auf `https://www.statsperform.com/artificial-intelligence`

Robertson, S., Bartlett, J. D. & Gastin, P. B. (2017). Red, Amber, or Green? Athlete Monitoring in Team Sport: The Need for Decision-Support Systems. *International Journal of Sports Physiology and Performance*, *12* (Suppl 2), S2–73-S2–79. doi: 10.1123/ijspp.2016-0541

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., ... Witvrouw, E. (2020). A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Medicine & Science in Sports & Exercise*, *52* (8), 1745-1751. doi: 10.1249/mss.0000000000002305

RStudio (Version 1.2.5019) [Computerprogramm]. (2019). RStudio. Zugriff am 07.11.2020 auf `https://download1.rstudio.org/desktop/bionic/amd64/rstudio-1.2.5019-amd64.deb`

Russel, S. & Norvig, P. (2012). *Künstliche Intelligenz* (3., aktualisierte Aufl.). München: Pearson.

R (Version 3.4.4) [Computerprogramm]. (o. J.). Launchpad. Zugriff am 06.11.2020 auf `https://launchpadlibrarian.net/366701586/r-base-core_3.4.4-1ubuntu1_amd64.deb`

Schlichtmeier, T. (2019). *Datenanalysen im Profi-Sport – Künstliche Intelligenz am Ball.* elektroniknet.de. Zugriff am 07.11.2020 auf `https://www.elektroniknet.de/elektronik/automation/kuenstliche-intelligenz-am-ball-164658.html`

Schnor, P. (2018). *Eine Künstliche Intelligenz auf Talentsuche im Profifußball.* Gründerszene. Zugriff am 07.11.2020 auf `https://www.gruenderszene.de/technologie/eine-kuenstliche-intelligenz-auf-talentsuche-im-profifussball?interstitial`

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27* (3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Silbernagel, S. (1991). *Taschenatlas der Physiologie* (4., überarbeitete Aufl.). Stuttgart: Georg Thieme.

Simple Logging Facade for Java (SLF4J) (Version 1.7.25) [Computerprogramm]. (o. J.). Quality Open Software. Zugriff am 14.12.2020 auf `https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.25/slf4j-api-1.7.25.jar`

Soccerlogic. (o. J.). Soccerlogic. Zugriff am 07.11.2020 auf `https://soccerlogic.com`

Soccerment. (o. J.). Soccerment. Zugriff am 07.11.2020 auf `https://soccerment.com/`

Soccerment Analytics. (o. J.). Soccerment. Zugriff am 07.11.2020 auf `https://analytics.soccerment.com`

StatsBomb is built by football experts. (o. J.). StatsBomb. Zugriff am 07.11.2020 auf `https://statsbomb.com/teams/`

Steinberg, E. (2020). PlantUML integration (Version 2.23.0) [Computerprogramm]. GitHub. Zugriff am 07.11.2020 auf `https://github.com/esteinberg/plantuml4idea/archive/master.zip`

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, *3* (17). doi: 10.1038/s41746-020-0221-y

TeX Live (Version 2017.20180305) [Computerprogramm]. (2018). Zugriff am 07.11.2020 auf `https://ubuntu.pkgs.org/18.04/ubuntu-universe-amd64/texlive-full_2017.20180305-1_all.deb.html`

Tomasits, J. & Haber, P. (2016). *Leistungsphysiologie* (5. Aufl.). Berlin [u.a.]: Springer.

Trabold, O., Wagner, S., Wicke, C., Scheuenstuhl, H., Hussain, M. Z., Rosen, N., ... Hunt, T. K. (2003). Lactate and oxygen constitute a fundamental regulatory mechanism in wound healing. *Wound Repair and Regeneration*, *11* (6), 504–509. doi: 10.1046/j.1524-475x.2003.11621.x

van der Zwaard, S., de Ruiter, C. J., Jaspers, R. T. & de Koning, J. J. (2019). Anthropometric Clusters of Competitive Cyclists and Their

Sprint and Endurance Performance. *Frontiers in Physiology*, *10*, 1276. doi: 10.3389/fphys.2019.01276

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4. Aufl.). o.O.: Springer.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Contributors, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261-272. doi: 10.1038/s41592-019-0686-2

Wahl, P., Bloch, W. & Mester, J. (2009). Moderne Betrachtungsweisen des Laktats: Laktat ein überschätztes und zugleich unterschätztes Molekül. *Schweizerische Zeitschrift für Sportmedizin und Sporttraumatologie*, *57* (3), 100-107.

Ward, P., Windt, J. & Kempton, T. (2019). Business Intelligence: How Sport Scientists Can Support Organisation Decision Making in Professional Sport. *International Journal of Sports Physiology and Performance*, *14* (4), 544-546. doi: 10.1123/ijspp.2018-0903

Webb, P., Syer, D., Long, J., Nicoll, S., Winch, R., AndyWilkinson, ... Simons, M. (2017). *Spring Boot Reference Guide.* Spring. Zugriff am 07.11.2020 auf `https://docs.spring.io/spring-boot/docs/1.5.8.RELEASE/reference/pdf/spring-boot-reference.pdf`

Welcome to Scriptella ETL Project. (o. J.). Scriptella. Zugriff am 06.11.2020 auf `http://scriptella.org/index.html`

What is Apatar Open Source Data Integration? (o. J.). Altoros. Zugriff am 07.11.2020 auf `https://www.altoros.com/blog/what-is-apatar-open-source-data-integration/`

What is BIRT? (o. J.). The Eclipse Foundation. Zugriff am 07.11.2020 auf `https://www.eclipse.org/birt/about/`

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2017). *Data Mining* (4. Aufl.). Boston [u.a.]: Elsevier, Morgan Kaufmann.

Yeung, J. (2018). *Analyzing, but not Predicting, this Year's World Cup.* SAP Community. Zugriff am 07.11.2020 auf `https://blogs.sap.com/2018/05/30/analyzing-but-not-predicting-this-years-fifa-world-cup/`

A Konfigurationen des prototypischen Systems

Das vorliegende Kapitel listet zunächst die Konfigurationen im JSON-Format für das Anlegen der Tabellen in der Basis- und Ableitungsdatenbank auf. Anschließend werden die Konfigurationen für die Datenintegration durch den CSV-Client im XML-Format aufgeführt.

A.1 Erstellen von Tabellen

Im Folgenden sind die Konfigurationen für das Anlegen der Tabellen in der Basis- und der Ableitungsdatenbank zu finden.

A.1.1 Basisdatenbank

Listing 8 auf S. 154 zeigt den Konfigurationsauschnitt der Stammdatentabelle.

```
{
  "tableName": "master_data",
  "tableColumns": [
    {
      "type": "bigint",
      "unique": "true",
      "name": "sd_id"
    },
    {
      "type": "varchar(10)",
      "name": "geschlecht"
    },
    {
      "type": "date",
      "name": "geburtsdatum"
    },
    {
      "type": "varchar(45)",
      "name": "sportart"
    },

    ...

  ]
}
```

Listing 8: Konfiguration der Tabelle master_data (Quelle: Eigene Darstellung)

Listing 9 auf S. 155-156 enthält Auszüge aus der Konfiguration für die Tabelle der ausdauerdiagnostischen Daten.

```
{
  "tableName": "endurance",
  "tableColumns": [
    {
      "type": "bigint",
      "unique": "true",
      "name": "sd_id"
    },
    {
      "type": "date",
      "unique": "true",
      "name": "untersuchungsdatum"
    },
    {
      "type": "int",
      "name": "alter_bei_untersuchung"
    },
    ...
        {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_2_4_zeit_km
        "
    },
    ...
    {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_2_8_zeit_km
        "
    },
    ...
    {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_3_2_zeit_km
        "
    },
    ...
    {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_3_6_zeit_km
        "
    },
    ...
    {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_4_0_zeit_km
        "
    },
    ...
    {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_4_4_zeit_km
        "
    },
    ...
    {
      "type": "varchar(10)",
      "name": "messwerte_test_geschwindigkeit_4_8_zeit_km
        "
    },
    ...
```

```
54      {
55        "type": "varchar(10)",
56        "name": "messwerte_test_geschwindigkeit_5_2_zeit_km
            "
57      },
58      ...
59      {
60        "type": "varchar(10)",
61        "name": "messwerte_test_geschwindigkeit_5_6_zeit_km
            "
62      },
63      ...
64      {
65        "type": "varchar(10)",
66        "name": "messwerte_test_geschwindigkeit_6_0_zeit_km
            "
67      },
68      ...
69      {
70        "type": "varchar(10)",
71        "name": "messwerte_test_geschwindigkeit_e3_zeit_km"
72      },
73      ...
74    ]
75  }
```

Listing 9: Konfiguration der Tabelle endurance (Quelle: Eigene Darstellung)

Ein Auszug aus der Konfiguration der Tabelle für die Daten der Laboratoriumsdiagnostik ist Listing 10 auf S. 156 zu entnehmen.

```
1  {
2    "tableName": "lab",
3    "tableColumns": [
4      {
5        "type": "bigint",
6        "unique": "true",
7        "name": "sd_id"
8      },
9      {
10       "type": "date",
11       "unique": "true",
12       "name": "untersuchungsdatum"
13     },
14
15     ...
16
17     {
18       "type": "float",
19       "name": "labor_haemoglobin"
20     },
21
22     ...
23
24   ]
25 }
```

Listing 10: Konfiguration der Tabelle lab (Quelle: Eigene Darstellung)

A.1.2 Ableitungsdatenbank

Die Konfiguration der Tabelle mit den aggregierten ausdauer- und laboratoriumsdiagnostischen Daten ist in Auszügen Listing 11 auf S. 157-158 zu entnehmen.

```
{
  "tableName": "endurance_lab",
  "tableColumns": [
    {
      "type": "bigint",
      "unique":"true",
      "name": "sd_id"
    },
    {
      "type": "date",
      "name": "untersuchungsdatum"
    },
    {
       "type":"varchar(10)",
       "name":"geschlecht"
    },
    {
       "type":"varchar(50)",
       "name":"sportart"
    },
    {
      "type": "date",
      "name": "geburtsdatum"
    },
    {
      "type": "int",
      "name": "`alter`"
    },
    {
      "type": "float",
      "name": "vo2_rel"
    },
    {
      "type": "float",
      "name": "rq"
    },
    {
      "type": "float",
      "name": "hf"
    },
    {
      "type": "float",
      "name": "laktat"
    },
    {
      "type": "int",
      "name": "zeit_bis_abbruch"
    },
    {
      "type": "float",
      "name": "geschw_laktat_4"
    },
```

```
    {
      "type": "float",
      "name": "erreichte_geschwindigkeit"
    },
    {
      "type": "float",
      "name": "hb"
    }
  ]
}
```

Listing 11: Konfiguration der Tabelle endurance_lab (Quelle: Eigene Darstellung)

Listing 12 auf S. 158-159 zeigt auszugsweise die Konfiguration für die Tabelle der aggregierten ausdauer- und laboratoriumsdiagnostischen Daten, ergänzt um die Cluster-Bezeichnung.

```
{
  "tableName": "endurance_lab_with_cluster_ids",
  "tableColumns": [
    {
      "type": "bigint",
      "unique":"true",
      "name": "sd_id"
    },
    {
      "type": "varchar(10)",
      "name": "cluster_id"
    },
    {
      "type": "date",
      "name": "untersuchungsdatum"
    },
    {
       "type":"varchar(10)",
       "name":"geschlecht"
    },
    {
       "type":"varchar(50)",
       "name":"sportart"
    },
    {
      "type": "date",
      "name": "geburtsdatum"
    },
    {
      "type": "int",
      "name": "`alter`"
    },
    {
      "type": "float",
      "name": "vo2_rel"
    },
    {
      "type": "float",
      "name": "rq"
```

```
    },
    {
      "type": "float",
      "name": "hf"
    },
    {
      "type": "float",
      "name": "laktat"
    },
    {
      "type": "int",
      "name": "zeit_bis_abbruch"
    },
    {
      "type": "float",
      "name": "geschw_laktat_4"
    },
    {
      "type": "float",
      "name": "erreichte_geschwindigkeit"
    },
    {
      "type": "float",
      "name": "hb"
    }
  ]
}
```

Listing 12: Konfiguration der Tabelle endurance_lab_with_cluster_ids (Quelle: Eigene Darstellung)

A.2 Datenintegration

Das vorliegende Unterkapitel enthält die Konfiguration des CSV-Clients für die Integration der verschiedenen Daten in die Basis- und Ableitungsdatenbank.

A.2.1 Basisdatenbank

Listing 13 auf S. 160 enthält die Konfiguration für die Integration der Stammdaten.

```
<config>
    <url>
        <protocol>https</protocol>
        <hostname>localhost</hostname>
        <port>2048</port>
        <path>/dataBulkUploadByRow</path>
    </url>
    <csv-file-path>$path/master_data.csv</csv-file-path>
    <tablename>master_data</tablename>
    <columnmappings>
    <mapping>
        <csvcolumnname>alter</csvcolumnname>
        <tablecolumnname>alter_bei_untersuchung</
           tablecolumnname>
    </mapping>
    <mapping>
        <csvcolumnname>SD-ID</csvcolumnname>
        <tablecolumnname>sd_id</tablecolumnname>
    </mapping>
    </columnmappings>
</config>
```

Listing 13: Konfiguration der Stammdatenintegration (Quelle: Eigene Darstellung)

Die Konfiguration für die Integration der Ausdauerdiagnostikdaten kann Listing 14 auf S. 160-161 entnommen werden.

```
<config>
    <url>
        <protocol>https</protocol>
        <hostname>localhost</hostname>
        <port>2048</port>
        <path>/dataBulkUploadByRow</path>
    </url>
    <csv-file-path>$path/endurance.csv</csv-file-path>
    <tablename>endurance</tablename>
    <columnmappings>
    <mapping>
        <csvcolumnname>alter</csvcolumnname>
```

```
13          <tablecolumnname>alter_bei_untersuchung</
                tablecolumnname>
14      </mapping>
15      </columnmappings>
16  </config>
```

Listing 14: Konfiguration der Integration von Ausdauerdiagnostikdaten (Quelle: Eigene Darstellung)

Die Integration der Laboratoriumsdiagnostikdaten wird entsprechend Listing 15 auf S. 161 konfiguriert.

```
1  <config>
2      <url>
3          <protocol>https</protocol>
4          <hostname>localhost</hostname>
5          <port>2048</port>
6          <path>/dataBulkUploadByRow</path>
7      </url>
8      <csv-file-path>$path/lab.csv</csv-file-path>
9      <tablename>lab</tablename>
10     <columnmappings>
11     <mapping>
12         <csvcolumnname>alter</csvcolumnname>
13         <tablecolumnname>alter_bei_untersuchung</
               tablecolumnname>
14     </mapping>
15     </columnmappings>
16 </config>
```

Listing 15: Konfiguration der Integration von Laboratoriumsdiagnostikdaten (Quelle: Eigene Darstellung)

A.2.2 Ableitungsdatenbank

Aus Listing 16 auf S. 161-162 kann die Integrationskonfiguration für die aggregierten Datensätze entnommen werden.

```
1  <config>
2      <url>
3          <protocol>https</protocol>
4          <hostname>localhost</hostname>
5          <port>2050</port>
6          <path>/dataBulkUploadByRow</path>
7      </url>
8      <csv-file-path>$path/endurance_lab.csv</csv-file-path
           >
9      <tablename>endurance_lab</tablename>
10     <columnmappings>
11             <mapping>
12             <csvcolumnname>alter</csvcolumnname>
13             <tablecolumnname>‘alter‘</tablecolumnname>
14         </mapping>
```

```
        <mapping>
            <csvcolumnname>VO2 rel</csvcolumnname>
            <tablecolumnname>vo2_rel</tablecolumnname>
        </mapping>
        <mapping>
            <csvcolumnname>Rq</csvcolumnname>
            <tablecolumnname>rq</tablecolumnname>
        </mapping>
        <mapping>
            <csvcolumnname>ges Dauer</csvcolumnname>
            <tablecolumnname>zeit_bis_abbruch</
               tablecolumnname>
        </mapping>
            <mapping>
            <csvcolumnname>Gesch Lak 4</csvcolumnname>
            <tablecolumnname>geschw_laktat_4</
               tablecolumnname>
        </mapping>
        </columnmappings>
</config>
```

Listing 16: Integrationskonfiguration der aggregierten Datensätze (Quelle: Eigene Darstellung)

Listing 17 auf S. 162-163 enthält die Integrationskonfiguration der modifizierten Datensätze.

```
<config>
    <url>
        <protocol>https</protocol>
        <hostname>localhost</hostname>
        <port>2048</port>
        <path>/dataBulkUploadByRow</path>
    </url>
    <csv-file-path>$path/endurance_lab__10_11__ward_2_39
       .0__with_cluster_ids.csv</csv-file-path>
    <tablename>endurance_lab_with_cluster_id</tablename>
    <columnmappings>
                <mapping>
            <columnname>Cluster-ID</columnname>
            <tablecolumnname>cluster_id</tablecolumnname>
        </mapping>
                <mapping>
            <columnname>alter</columnname>
            <tablecolumnname>`alter`</tablecolumnname>
        </mapping>
        <mapping>
            <columnname>VO2 rel</columnname>
            <tablecolumnname>vo2_rel</tablecolumnname>
        </mapping>
        <mapping>
            <columnname>Rq</columnname>
            <tablecolumnname>rq</tablecolumnname>
        </mapping>
        <mapping>
            <columnname>ges Dauer</columnname>
            <tablecolumnname>zeit_bis_abbruch</
               tablecolumnname>
```

```
30          </mapping>
31                  <mapping>
32              <columnname>Gesch Lak 4</columnname>
33              <tablecolumnname>geschw_laktat_4</
                    tablecolumnname>
34          </mapping>
35                  <mapping>
36              <columnname>er Geschw</columnname>
37              <tablecolumnname>erreichte_geschwindigkeit</
                    tablecolumnname>
38          </mapping>
39      </columnmappings>
40  </config>
```

Listing 17: Integrationskonfiguration der modifizierten Datensätze (Quelle: Eigene Darstellung)

B Physiologische Parameter in Bezug auf die Zeit bis zum Abbruch

Innerhalb dieses Kapitels werden die physiologischen Parameter rVO_{2peak}, RQ_{peak}, Hf_{max}, Lak_{peak} sowie Hb und der Leistungsparameter tlim mit Hilfe von Streudiagrammen graphisch in Bezug zueinander gebracht und kurz beschrieben.

B.1 Relative maximale Sauerstoffaufnahme (peak)

In Abb. 39 auf S. 164 ist ein Streudiagramm mit den beiden Parametern rVO_{2peak} und tlim zu sehen. Auf der x-Achse sind die Werte der rVO_{2peak}

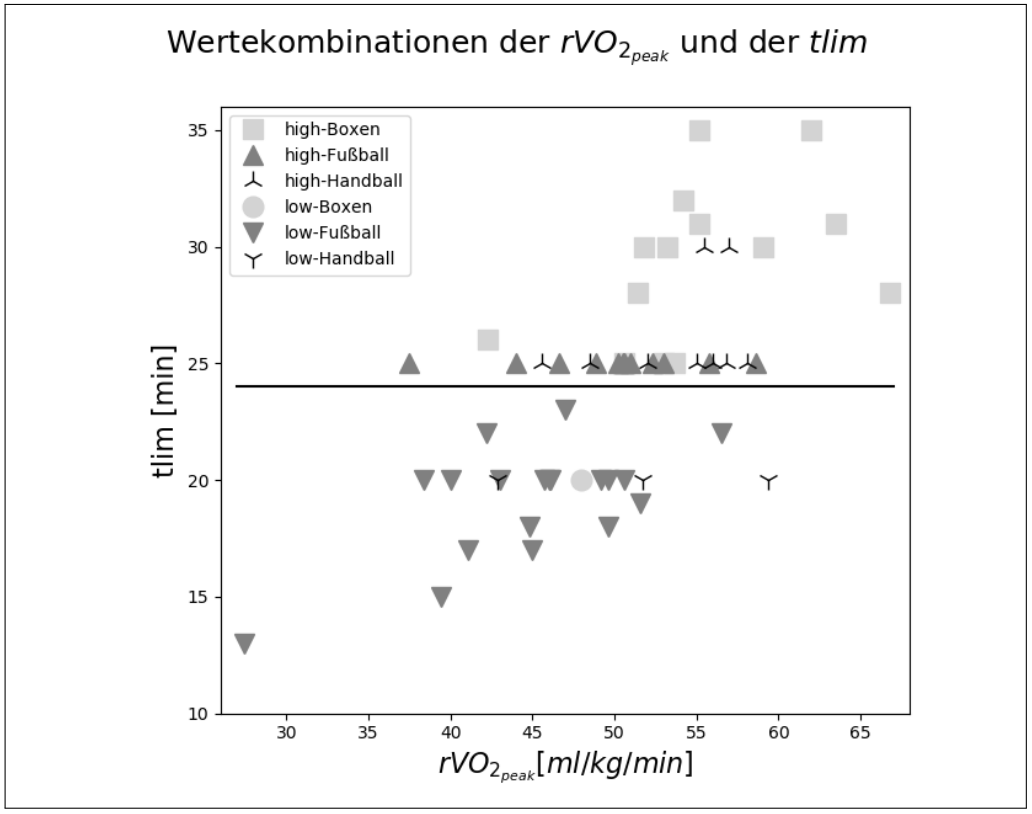

Abbildung 39: Wertekombinationen der rVO_{2peak} und der tlim (Quelle: Eigene Darstellung)

mit einer fünfstufigen Skala eingezeichnet. Die Skala reicht von 30 bis 65 ml/kg/min. Die y-Achse zeigt die tlim, die ebenfalls eine fünfstufige Skala aufweist. Innerhalb der Skala sind Werte zwischen 10 und 35 min verzeichnet. Jedes Individuum ist entsprechend seiner Gruppenzugehörigkeit durch ein Symbol gekennzeichnet. Individuen der Gruppe high-Boxen sind durch ein hellgraues Quadrat, die der Gruppe high-Fußball durch ein dunkelgraues Dreieck mit nach oben zeigender Spitze und die der Gruppe high-Handball durch einen dreizackigen Stern mit nach oben gerichteter Spitze gekennzeichnet. Für Individuen der Gruppe low-Boxen existiert ein hellgrauer Kreis, für die der Gruppe low-Fußball ein nach unten gerichtetes dunkelgraues Dreieck und für die Individuen der Gruppe low-Handball ein nach unten gerichteter dreizackiger Stern. Eine schwarze horizontale Linie trennt die Wertekombinationen clusterweise auf Höhe von 24 min auf der y-Achse und einer Strecke von 27 bis 67 ml/kg/min auf der x-Achse.
Ausprägungen ab ca 28 min tlim weisen Werte für die $rVO_{2_{peak}}$ ab 50 ml/min/kg auf. Diese Wertekombinationen sind hauptsächlich für Individuen der Gruppe high-Boxen zu finden. Zwei dieser Kombinationen können für Individuen der Gruppe high-Handball verzeichnet werden. Bei Werten von 25 und 26 min bei der tlim sind Werte für die $rVO_{2_{peak}}$ zwischen ca 35 und 60 ml/kg/min für alle Gruppen aus Cluster high zu beobachten. Mit einer Ausnahme von unter 15 min bei der tlim und unter 30 ml/kg/min bei der $rVO_{2_{peak}}$ liegen die Werte aus Cluster low alle zwischen einer tlim von 15 und 24 min sowie einer $rVO_{2_{peak}}$ zwischen 35 und 60 ml/kg/min.

B.2 Respiratorischer Quotient (peak)

Dem Streudiagramm in Abb. 40 auf S. 166 können die Wertekombinationen für die Parameter RQ_{peak} und tlim entnommen werden. Die x-Achse enthält den Parameter RQ_{peak} mit einem Wertebereich zwischen 0.95 und 1.3 sowie einer Skalierung von 0.05. Auf der y-Achse ist der Parameter tlim mit einer fünfstufigen Skalierung bei einem Wertebereich zwischen 10 und 35 min zu sehen. Die Individuen sind entsprechend ihrer Gruppenzugehörig-

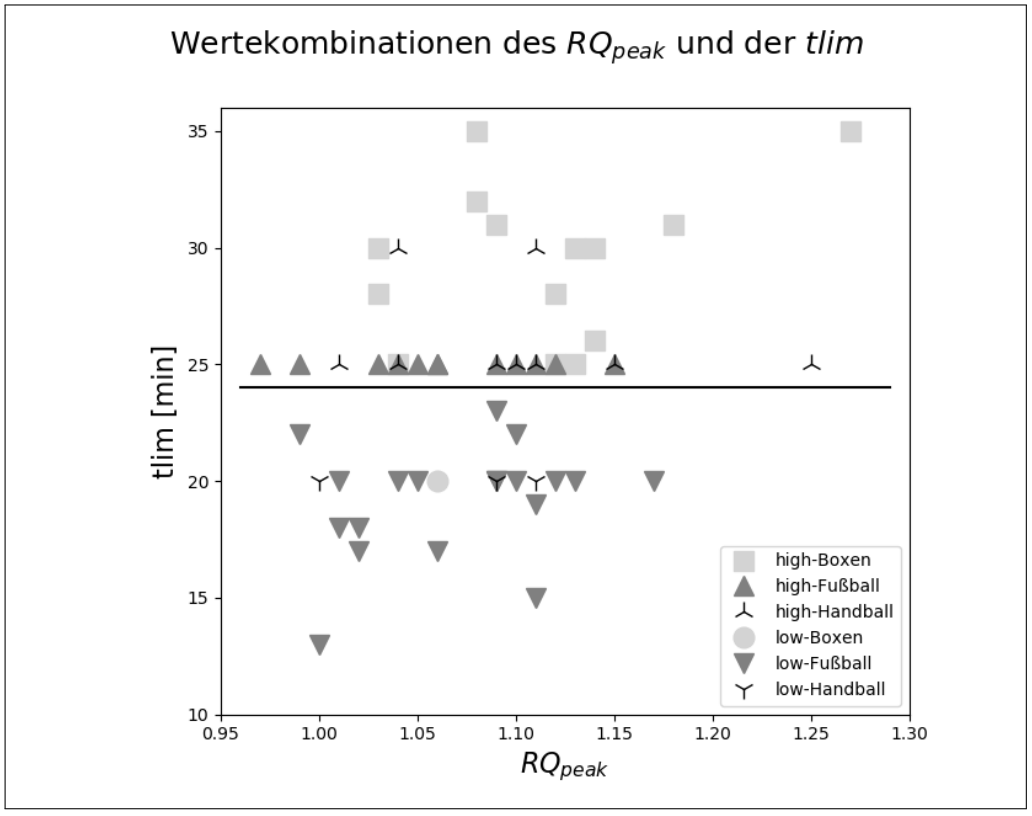

Abbildung 40: Wertekombinationen des RQ_{peak} und der tlim (Quelle: Eigene Darstellung)

keit durch ein entsprechendes Symbol gekennzeichnet. Individuen der Gruppe high-Boxen sind durch ein hellgraues Quadrat, die der Gruppe high-Fußball durch ein dunkelgraues Dreieck mit nach oben zeigender Spitze und die der Gruppe high-Handball durch einen dreizackigen Stern mit nach oben gerichteter Spitze gekennzeichnet. Für Individuen der Gruppe low-Boxen existiert ein hellgrauer Kreis, für die der Gruppe low-Fußball ein nach unten gerichtetes dunkelgraues Dreieck und für die Individuen der Gruppe low-Handball ein nach unten gerichteter dreizackiger Stern. Die Wertekombinationen der beiden Cluster sind durch eine horizontale schwarze Linie auf der Höhe von 24 min bei der tlim sowie einem Wertebereich zwischen 0.96 und 1.29 beim RQ_{peak} getrennt.
Zwischen Werten von 0.95 und 1.2 für den RQ_{peak} sind verschiedenste Wertekombinationen vertreten. Hierbei sind die verschiedenen Wertekombinationen über alle Gruppen hinweg zu finden. Die beiden höchsten Werte mit einem RQ_{peak} von über 1.2 sind bei Individuen der Gruppen high-Boxen und high-Handball zu finden.

B.3 Maximale Herzfrequenz

Der Abbildung 41 auf S. 168 kann ein Streudiagramm mit Wertekombinationen für die Parameter Hf_{max} und tlim entnommen werden. Auf der x-Achse des Diagramms ist der Parameter Hf_{max} mit einer zehnstufigen Skala eingezeichnet. Die Werte der Skala liegen dabei zwischen 150 und 220 S/min. Die y-Achse enthält den Parameter tlim mit einer fünfstufigen Skala und einem Wertebereich zwischen 10 und 35 min. Jedes Individuum ist entsprechend seiner Gruppenzugehörigkeit durch ein Symbol gekennzeichnet. Individuen der Gruppe high-Boxen sind durch ein hellgraues Quadrat, die der Gruppe high-Fußball durch ein dunkelgraues Dreieck mit nach oben zeigender Spitze und die der Gruppe high-Handball durch einen dreizackigen Stern mit nach oben gerichteter Spitze gekennzeichnet. Für Individuen der Gruppe low-Boxen existiert ein hellgrauer Kreis, für die der Gruppe low-Fußball ein nach unten gerichtetes dunkelgraues Dreieck und für die Individuen der Gruppe low-Handball ein nach unten gerichteter dreizackiger Stern. Die Wertekom-

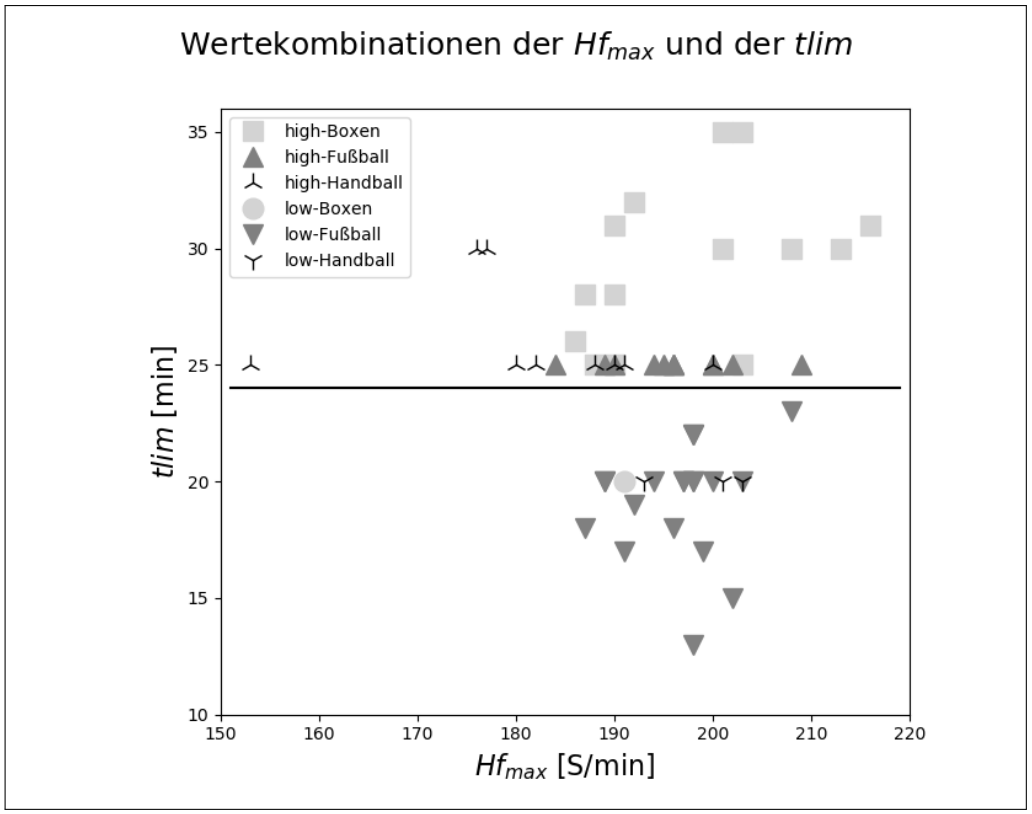

Abbildung 41: Wertekombinationen der $\mathrm{Hf_{max}}$ und der tlim (Quelle: Eigene Darstellung)

binationen sind clusterweise durch eine schwarze Linie getrennt. Diese liegt bei einem Wert von 24 min bei der tlim und erstreckt sich bei der Hf_{max} zwischen 151 und 219 S/min.
Überwiegend sind in dem Streudiagramm Werte für die $Hf_{max} \geq 184$ S/min für alle Gruppen zu verzeichnen. Dabei treten unterschiedliche Wertekombinationen unabhängig von der Höhe der Hf_{max} auf. Es existieren darüber hinaus fünf Wertekombinationen mit einer Hf_{max} von unter 184 S/min. Diese Kombinationen sind alle Individuen aus der Gruppe high-Handball zuzuordnen. Die beiden höchsten Werte für die Hf_{max} mit über 210 S/min stammen von Individuen der Gruppe high-Boxen.

B.4 Blutlaktatkonzentration (peak)

Abb. 42 auf S. 170 zeigt ein Streudiagramm mit den beiden Parametern Lak_{peak} und tlim. Der Parameter Lak_{peak} ist auf der x-Achse, der Parameter tlim auf der y-Achse eingezeichnet. Die x-Achse enthält eine einstufige Skala mit Werten zwischen 4 und 11 mmol/l, die y-Achse eine fünfstufige Skala mit Werten zwischen 10 und 35 min. Die einzelnen Wertekombinationen der jeweiligen Individuen sind entsprechend der Gruppenzugehörigkeit durch verschiedene Symbole gekennzeichnet. Ein hellgraues Quadrat symbolisiert die Gruppe high-Boxen, ein dunkelgraues Dreieck mit der Spitze nach oben die Gruppe high-Fußball und ein dreizackiger Stern mit einer Spitze nach oben die Gruppe high-Handball. Die Gruppen low-Boxen, low-Fußball und low-Handball werden durch einen hellgrauen Kreis, ein dunkelgraues Dreieck mit der Spitze nach unten sowie einen dreizackigen Stern mit einer Spitze nach unten dargestellt. Die Wertekombinationen sind clusterweise durch eine schwarze horizontale Linie auf der Höhe von 24 min auf der y-Achse und einer Strecke von 4.1 bis 11.4 mmol/l auf der x-Achse voneinander getrennt. Alle Wertepaare > 25 min für tlim sind – bis auf zwei Ausnahmen der Gruppe high-Handball – der Gruppe high-Boxen zugeordnet. Die Werte für die Lak_{peak} liegen dabei innerhalb eines Intervalls zwischen ungefähr 5.6 und 11.1 mmol/l. Die Gruppen high-Boxen, high-Fußball und high-Handball weisen Wertekombinationen bei der tlim = 25 min innerhalb des Intervalls 4.3

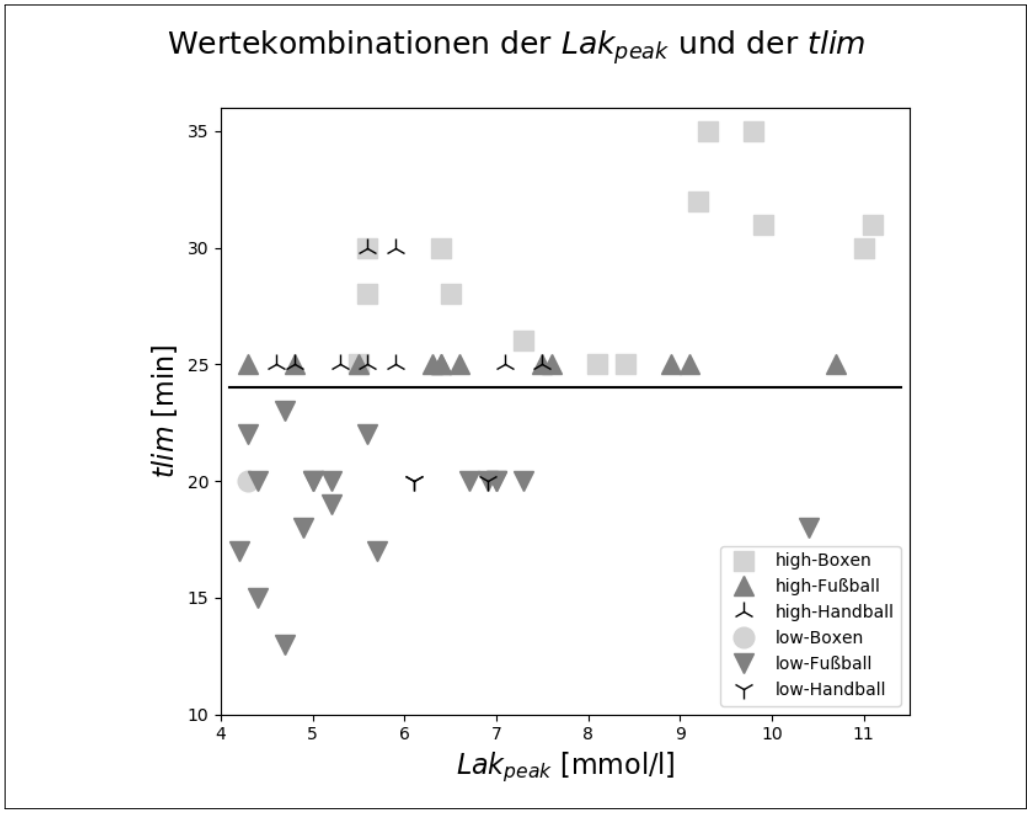

Abbildung 42: Wertekombinationen der $\mathrm{Lak}_{\mathrm{peak}}$ und der tlim (Quelle: Eigene Darstellung)

und 10.7 mmol/l für den Parameter Lak_{peak} auf.
Unterhalb von 25 min für die tlim sind Wertekombinationen in einem Intervall von 4.2 bis 10.4 mmol/l für die Lak_{peak} für die Gruppen low-Boxen, low-Fußball und low-Handball vorhanden.

B.5 Hämoglobin-Wert

Der Abbildung 43 auf S. 171 kann ein Streudiagramm mit individuellen Wertekombinationen für den Hb und die tlim entnommen werden. Auf der x-

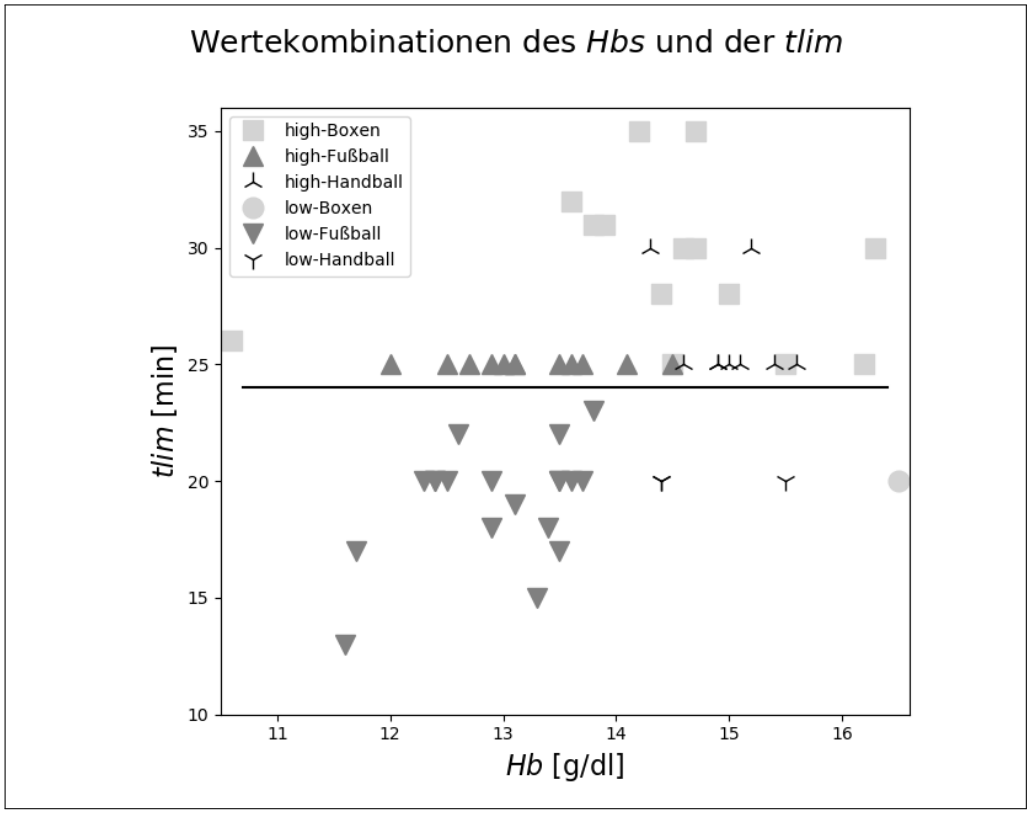

Abbildung 43: Wertekombinationen des Hbs und der tlim (Quelle: Eigene Darstellung)

Achse des Plots ist der Parameter Hb verzeichnet. Die Skala der x-Achse ist einstufig und umfasst Werte zwischen 11 und 16 g/dl. Die y-Achse weist den Parameter tlim in einer fünfstufigen Skala mit Werten zwischen 10 und 35 min auf. Die einzelnen Wertekombinationen sind nach Gruppen durch un-

terschiedliche Symbole gekennzeichnet. Ein hellgraues Quadrat symbolisiert die Gruppe high-Boxen, ein dunkelgraues Dreieck mit Spitze nach oben die Gruppe high-Fußball und ein dreizackiger Stern mit einer Spitze nach oben die Gruppe high-Handball. Die Gruppen low-Boxen, low-Fußball und low-Handball werden durch einen hellgrauen Kreis, ein dunkelgraues Dreieck mit Spitze nach unten sowie einen dreizackigen Stern mit einer Spitze nach unten dargestellt. Eine schwarze Linie trennt die Wertekombinationen nach Clustern bei der tlim in der Höhe von 24 min und Werten des Hbs zwischen 10.7 und 16.4 g/dl.
Die Wertekombinationen tlim $<$ 25 min weisen Werte für den Hb zwischen 11.6 und 16.3 g/dl auf. Bei einem Wert $>$ 24 min bei der tlim sind Werte zwischen 10.6 und 16.3 g/dl für den Hb zu finden. Des Weiteren sind Kombinationen mit Werten für den Hb $\leq$ 13.6 g/dl überwiegend den Gruppen high- und low-Fußball zuzuordnen, die Kombinationen mit Werten darüber hauptsächlich Individuen der Gruppen high- und low-Boxen und Handball.

C Stabilitätsanalyse des Clusterings

Für die Stabilitätsanalyse des in dieser Arbeit vorgenommenen Clusterings[112] wurden 9 der 58 und damit jeder sechste der vorhandenen Datensätze entfernt.

Abb. 44 auf S. 173 enthält die graphische Darstellung des durch die Stabilitätsanalyse enstandenen Dendogramms. Dieses zeigt auf der x-Achse die in das Clustering eingeflossenen Objekte, welche durch eine fortlaufende Integerzahl von 1 bis 49 innerhalb des Diagramms repräsentiert werden. Auf der y-Achse sind die Distanzen dargestellt. Die Skalierung ist fünfstufig und umfasst die Werte von 0 bis 40. Die Linie des cut-off ist bei der Distanz von

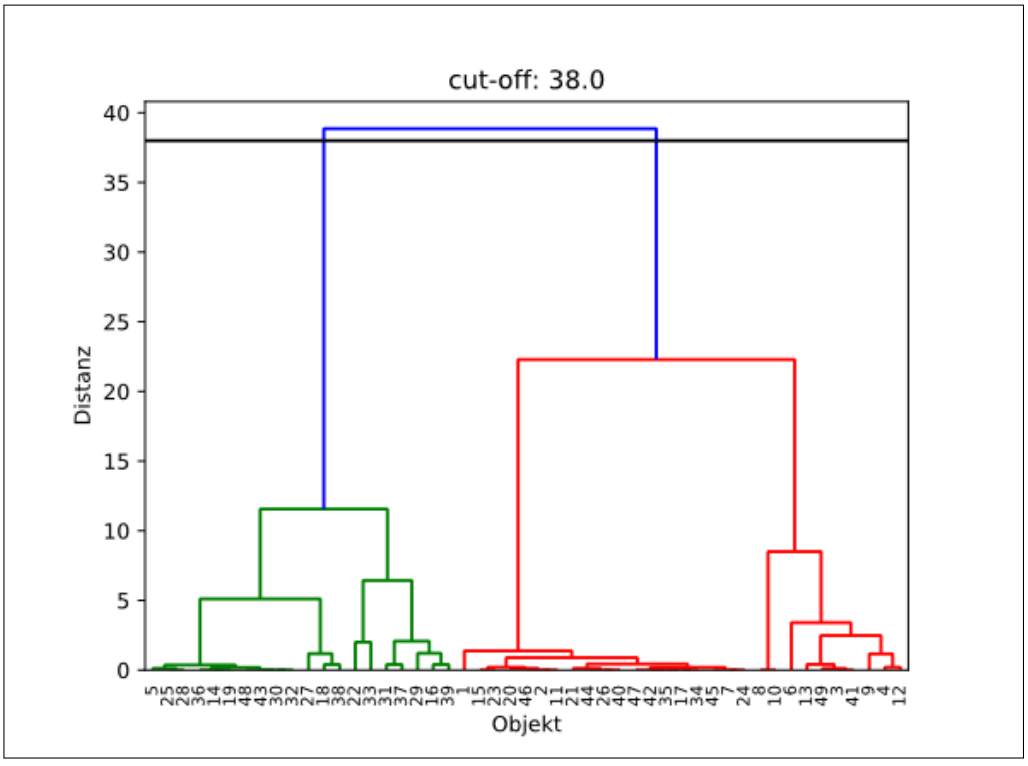

Abbildung 44: Graphische Darstellung des Dendogramms aus der Stabilitätsanalyse (Quelle: Eigene Darstellung)

38 zu sehen. Diese Linie ist bei der größten Distanz zwischen zwei Clustern

[112] Siehe Kapitel 5.1 ab S. 80

eingezeichnet. Durch die Linie werden zwei vertikale Linien und somit zwei Cluster gekreuzt, so dass bei dem durchgeführten Clustering ebenfalls von zwei Clustern ausgegangen werden kann. Im Folgenden wird Cluster 1 wiederum als Cluster low, Cluster 2 als Cluster high bezeichnet.

Tabelle 17 auf S. 174 gibt Aufschluss über die MWe von Cluster low. Die Tabelle enthält drei Spalten, je eine mit den Parametern, den Einheiten des jeweiligen Parameters sowie dem jeweiligen MW. Die *tlim* weist laut der Tabelle einen Wert von 19.2 min auf. Für die *V4* ist ein Wert von 3.33 m/s zu verzeichnen. Für die $rVO_{2_{peak}}$ beträgt der MW 46.7 ml/kg/min. Der RQ_{peak} kann mit 1.07 beziffert werden. Die Hf_{max} weist einen Wert von 196 S/min auf. Die Lak_{peak} ist mit 5.82 mmol/l bestimmt. Die Individuen aus Cluster low weisen im Mittel einen Wert von 13.3 g/dl für den *Hb* auf.

Parameter	Einheit	MW
tlim	min	19.2
V4	m/s	3.33
$rVO_{2_{peak}}$	ml/kg/min	46.7
RQ_{peak}	-	1.07
Hf_{max}	S/min	196
Lak_{peak}	mmol/l	5.82
Hb	g/dl	13.3

Tabelle 17: Parametermittelwerte von Cluster low der Stabilitätsanalyse

Tabelle 18 auf S. 175 enthält die Parameter, die jeweiligen Einheiten sowie die jeweiligen MWe von Cluster high. Die Individuen aus Cluster high weisen im Mittel für die *tlim* einen Wert von 27.2 min auf. Der Parameter *V4* besitzt einen Wert von 3.81 m/s, die $rVO_{2_{peak}}$ einen Wert von 53.43 ml/kg/min. 1.09 beträgt der Wert des RQ_{peak}. Für den Parameter Hf_{max} liegt ein Wert von 193 S/min vor. Auf 7.24 mmol/l ist der Wert für den Parameter Lak_{peak} zu beziffern. Schließlich beträgt der Mittelwert für den *Hb* 14.2 g/dl.

Abb. 45 auf S. 175 zeigt ein Parallelkoordinatendiagramm, in dem die Mittelwerte der beiden Cluster für die einzelnen untersuchten Parameter einge-

Parameter	Einheit	MW
tlim	min	27.2
V4	m/s	3.81
$rVO2_{peak}$	ml/kg/min	53.43
RQ_{peak}	-	1.09
Hf_{max}	S/min	193
Lak_{peak}	mmol/l	7.24
Hb	g/dl	14.2

Tabelle 18: Parametermittelwerte von Cluster high der Stabilitätsanalyse

zeichnet sind. Mit Hilfe des Diagramms werden im Folgenden die Wertediffe-

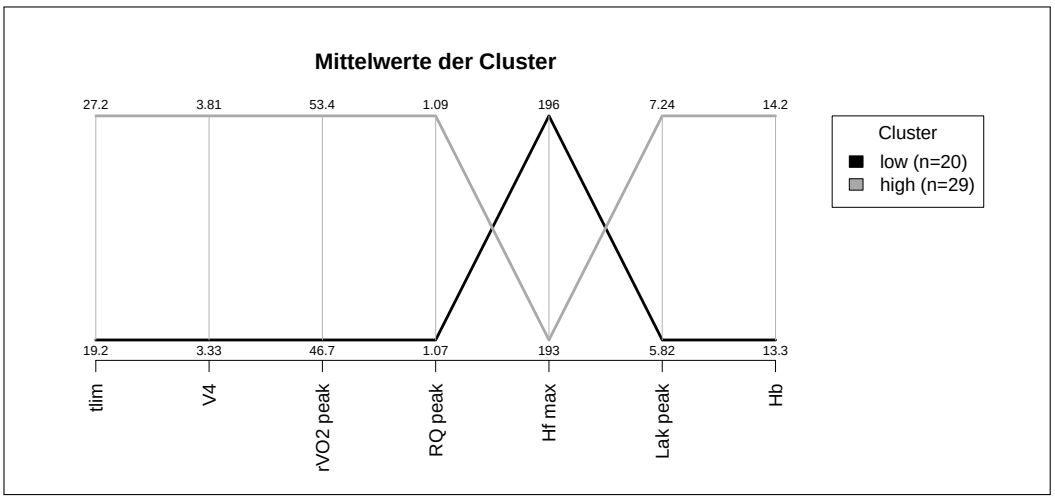

Abbildung 45: Parametermittelwerte der Cluster aus der Stabilitätsanalyse (Quelle: Eigene Darstellung)

renzen bei den verschiedenen Parametern zwischen Cluster low und Cluster high veranschaulicht und beschrieben. Die x-Achse des Diagramms enthält die untersuchten Parameter *tlim*, *V4*, $rVO_{2_{peak}}$, RQ_{peak}, Hf_{max}, Lak_{peak} und *Hb*. Auf den jeweiligen y-Achsen ist die entsprechende Skalierung der Parameter enthalten. Die Mittelwerte von Cluster low sind schwarz, die von Cluster high grau markiert. Cluster low weist 20 Individuen auf, Cluster high 29. Dem Diagramm ist zu entnehmen, dass die Werteausprägungen von Cluster low für die einzelnen Parameter mit Ausnahme des Parameters Hf_{max}

im Mittel niedriger sind als die von Cluster high.
Im Folgenden werden die jeweiligen Differenzen für die einzelnen Parameter beschrieben. Bei der tlim liegt eine Differenz von 8 min vor. Für den Parameter V4 kann eine Differenz von 0.48 festgestellt werden. Die $rVO_{2_{peak}}$ weist eine Differenz von 6.73 ml/kg/min auf. Der RQ_{peak} besitzt eine Differenz von nur 0.02. Für die Hf_{max} ist eine Differenz von 3 S/min zu verzeichnen. Die Hf_{max} ist der einzige Parameter, bei dem die Individuen von Cluster high im Mittel einen niedrigeren Wert als die von Cluster low aufweisen. Des Weiteren weist der Parameter Lak_{peak} eine Differenz von 1.42 mmol/l zwischen den Mittelwerten von Cluster low und Cluster high auf. Schließlich ist für den Parameter Hb eine Differenz von 0.9 g/dl zu verzeichnen.

Im Folgenden werden die Ergebnisse des Clusterings und die seiner Stabilitätsanalyse beschrieben und kurz miteinander verglichen.
Zunächst wird die prozentuale Verteilung der Individuen auf die Cluster low und high sowohl im Clustering selber als auch in der Stabilitätsanalyse betrachtet. Cluster low weist in der Stabilitätsanalyse ca. 41 Prozent der Individuen auf. Im Vergleich dazu weist Cluster low des Clusterings ca. 40 Prozent der Individuen auf. Cluster high sind in der Stabilitätsanalyse ca. 59 Prozent der Individuen zugeordnet, im Clustering ca. 60 Prozent. Somit kann die prozentuale Verteilung der Individuen auf die beiden Cluster sowohl beim Clustering als auch bei der Stabilitätsanalyse als sehr ähnlich betrachet werden, da sie sich nur um 1 Prozent unterscheidet.
Die beiden Parameter der Leistungsdiagnostik tlim und V4 stellen die Parameter dar, über die das Clustering stattgefunden hat. Die MWe der beiden Parameter unterscheiden sich für das leistungsschwächere Cluster low um 0.1 min (tlim) beziehungsweise um 0.01 m/s (V4). Für Cluster high liegen die Unterschiede in den Paramterausprägungen im Mittel bei 0.2 min (tlim) sowie 0 m/s (V4).
Die folgenden Differenzen der physiologischen Parameterausprägungen können für Cluster low zwischen dem Clustering und der Stabilitätsanalyse festgestellt werden. Die Differenz für die $rVO_{2_{peak}}$ beträgt 0.8 ml/kg/min, die für

den RQ_{peak} 0. Bei der Hf_{max} tritt eine Differenz von 1 S/min auf. Bei der Lak_{peak} beträgt diese 0.08 mmol/l. Die Mittelwerte für den Parameter Hb unterscheiden sich nicht.
Bei Cluster high liegen im Mittel ebenfalls sehr niedrige Differenzen zwischen den Ausprägungen der physiologischen Parameter beim Clustering und der Stabilitätsanalyse vor. So beträgt die Differenz bei der rVO_{2peak} 0.43 ml/kg/min, während beim Parameter RQ_{peak} sogar keine Differenz vorhanden ist. Auch die Hf_{max} unterscheidet sich nicht. Die Differenz der Lak_{peak} beläuft sich auf 0.09 mmol/l. Beim Hb ist keine Differenz zu verzeichnen.

Zusammenfassend kann festgehalten werden, dass zwischen den Mittelwerten des Clusterings und denen der Stabilitätsanalyse nur geringfügige Abweichungen existieren. Daher kann von stabilen Clustern ausgegangen werden.

Zusammenfassung

Aktive verschieben auf der Jagd nach neuen Rekorden – unterstützt durch Erkenntnisse aus der Sportwissenschaft – immer weiter die Grenzen der menschlichen Leistungsfähigkeit. Die Erkenntnisse der Sportwissenschaft fußen auf der Erhebung und Analyse verschiedenster Daten.
Dabei werden personalisierte Ansätze, wie sie bereits in der Medizin Anwendung finden, auch in der Sportwissenschaft immer wichtiger, um bei Entscheidungen im Bereich der Diagnostik und des Trainings unterstützen zu können.
Um solche Ansätze umsetzen zu können, werden ganz allgemein Prozesse zur Datenorganisation, Datenanalyse auf der Basis von künstlicher Intelligenz und Prozesse für die Kommunikation der aus der Analyse gewonnenen Erkenntnisse gefordert. Im Bereich der Ausdauerdiagnostik und des Ausdauertrainings existieren aktuell in der sportwissenschaftlichen Forschung wenige solcher allgemeinen, umgesetzten Ansätze.

Basierend auf den oben genannten Forderungen wurden innerhalb der vorliegenden Arbeit als übergeordneter Prozess der Data-Warehouse-Prozess sowie als Prozess für die Datenanalyse der CRoss Industry Standard Process for Data Mining (CRISP-DM) umgesetzt. Dazu wurde ein leichtgewichtiges Server-/Client-System mit einer JSON-API und einer relationalen Datenbank für die Datenorganisation entwickelt, um die hohen Kosten für die Anschaffung und Unterhaltung von großen Business-Intelligence-Systemen zu vermeiden. Basierend auf einem zweistufigen Machine-Learning-Modell wurde durch ein Clustering eine Einteilung von Individuen in zwei Leistungscluster vorgenommen. Für die beiden Leistungscluster wurden anschließend mit Hilfe eines Decision Trees Regeln gefunden, um Teile der physiologischen Strukturen beziehungsweise deren Ausprägungen aufzuzeigen.

Das erstellte Modell ermöglicht Einblicke in die physiologischen Strukturen von Leistungsclustern im Ausdauerbereich. Somit könnte das erstellte Modell das Betreuungspersonal von Aktiven im Leistungssport sowohl in der Dia-

gnostik als auch bei der Planung des Ausdauertrainings unterstützen. Die Anwendbarkeit sowie der Nutzen des Modells in der täglichen Praxis ist in zukünftigen Projekten zu evaluieren.

Abstract

Athletes chasing after new records – backed by the insights of sports science – outperform the boarders of humane capability. Those insights are based on the collection and analysis of diverse data.
In the years to come in sports science – similar to the advances in medical science – personalized approaches to support decision-making in the field of endurance diagnostics and endurance training are getting more and more important.
To follow such personalized approaches, processes of data organisation and data analyses with artificial intelligence as well as the communication of these insights have to be implemented.
At present, the use of such common approaches is scarce.

Based on the above-mentioned requirements the data warehouse process was implemented as the higher-ordered process, the subprocess for the data analysis by the CRoss Industry Standard Process for Data Mining (CRISP-DM). For this purpose a server-client-system with a JSON-API and an underlying relational database was developed to avoid the high costs of aquirement and maintenance of heavy business intelligence systems. Based on a two-step machine learning model a division of athletes into two endurance performance clusters was accomplished by using a hierarchical clustering algorithm. Afterwards rules were found by using a decision tree to identify parts of the physiological structures respectively their characteristics within the clusters.

Therefore the built model gives insights in the physiological structures of endurance performance clusters. Thus the model could be used to support trainers in endurance diagnostics and in planning endurance training. Future work has to be done to test the model in the daily diagnostic and training routine.

Danksagung

An dieser Stelle bedanke ich mich ganz herzlich bei meinem Doktorvater Herrn Univ.-Prof. Dr. Dr. hc. mult. Joachim Mester für die Möglichkeit der Promotion an der Deutschen Sporthochschule Köln sowie die Spezifizierung des Themas und das zur Verfügung stellen der in dieser Arbeit untersuchten Daten. Mein besonderer Dank gilt auch für sein offenes Ohr und die vielen Gespräche mit ihm, die mich sowohl fachlich als auch in meiner persönlichen Entwicklung sehr weiter gebracht haben. Ohne ihn wäre die vorliegende Arbeit in dieser Form nicht möglich gewesen.

Ebenso danke ich Frau Prof. Dr. Heide Faeskorn-Woyke sowohl für die Betreuung und ihren fachlich zielführenden Rat als auch für ihren unermüdlichen Einsatz innerhalb der Kooperation zwischen der Technischen Hochschule Köln und der Deutschen Sporthochschule Köln.

Des Weiteren danke ich Frau Prof. Dr. Edda Leopold posthum, die mir insbesondere im Hinblick auf das zielgerichtete Vorgehen bei der Anwendung von maschinellem Lernen mit ihrem Rat sehr zur Seite gestanden hat.

Und nicht zuletzt danke ich meinen ehemalige Kollegen des Instituts für Trainingswissenschaft und Sportinformatik. Neben dem häufigen fachlichen Austausch in Bezug auf sportwissenschaftliche Themen gilt mein Dank auch für die Erhebung der spiroergometrischen und labordiagnostischen Daten, ohne welche die vorliegende Arbeit nicht existieren würde.