Informatics

Visuelle Analyse der Genauigkeit von Vorhersagen für ATP-Matches

BACHELORARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science

im Rahmen des Studiums

Medieninformatik und Visual Computing

eingereicht von

Manuel Leutschacher
Matrikelnummer 11822449

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Zweitbetreuung: Dipl.-Ing. Dr. Johanna Schmidt

Wien, 22. Juni 2022

Manuel Leutschacher | Eduard Gröller

Informatics

Visual Analysis of the Prediction of ATP-Matches

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Media Informatics and Visual Computing

by

Manuel Leutschacher
Registration Number 11822449

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Second advisor: Dipl.-Ing. Dr. Johanna Schmidt

Vienna, 22nd June, 2022

Manuel Leutschacher | Eduard Gröller

Erklärung zur Verfassung der Arbeit

Manuel Leutschacher

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 22. Juni 2022

Manuel Leutschacher

Kurzfassung

Diese Arbeit befasst sich mit dem Thema der Vorhersage des Siegers eines Tennis-Matches auf der ATP-Tour. Zuerst wird in einer kurzen Einleitung die Relevanz und Forschung zu Machine Learning generell hervorgehoben und besprochen, wobei zuerst allgemeine und dann, spezifischer auf diese Arbeit zugeschnittene, Bereiche thematisiert werden.

Dann werden in einer ausführlicheren Literaturrecherche für dieses Thema interessante Arbeiten zusammengetragen, die den aktuellen Stand in der Forschung und auch am Markt widerspiegeln. Es wird auf Vorhersage-Algorithmen und aktuelle Ergebnisse der Vorhersageforschung eingegangen, dann werden speziell nochmals Ergebnisvorhersagen im Sportsektor hervorgehoben. Als sehr relevanter und lukrativer Einsatzbereich von Vorhersagen ist auch der Vorgehensweise der Wettanbieter/innen und Buchmacher/innen ein Kapitel gewidmet. Dies gilt ebenfalls für die Visualisierung von Daten im Zusammenhang mit Machine Learning, Frameworks und Libraries, um Machine Learning umsetzen zu können, und auch zukünftige Herausforderungen in diesem Bereich.

Dann, im praktischen Teil der Arbeit, wird eine Applikation vorgestellt, die die Vorhersageleistungen von mehreren verschiedenen neuronalen Netzen zur Gewinner Vorhersage von Tennis-Matches gegenüberstellt. Dafür wurde zuerst ein Datenset mit über 200.000 Einträgen und über 200 Variablen auf vier neuronale Netze aufgeteilt, die dann in Tensorflow erstellt und trainiert wurden. Die Modelle wurden zunächst anhand der Matches von drei Spielern getestet und die Ergebnisse dann in Streudiagrammen visualisiert. Es wird durch diese Vorgehensweise ersichtlich, welche Variablen, die vor dem Matchbeginn der Öffentlichkeit zur Verfügung stehen, wie relevant für den Ausgang eines Matches sind. Aber auch welches Set an Daten sich am besten dazu eignet, eine Vorhersage über den Gewinner zu tätigen und auch, ob es Unterschiede in den Vorhersagen im Vergleich dieser drei Spieler gibt.

Abstract

This Paper is centered around the topic of predicting the winner of a Tennis-Match on the ATP-Tour. Firstly there is a short introduction about the relevance and research of Machine Learning. At first the topic will be discussed broadly, then more and more narrowly to match the goal of the thesis. Next a research for important literature will be carried out, to present the state of the art in the field of Machine Learning based predictions, in science and on the market. Prediction algorithms will be discussed, with an emphasis on outcome prediction in different sports. As an important and high revenue usecase of prediction, the methods of betting companies and bookmakers also get a chapter to be discussed. Same applies to the visualization of data in conjunction to Machine Learning, related Frameworks and Libraries and lastly also future challenges of applying Machine Learning

In the practical part of the thesis, an application will be talked about, that compares the performance of different neural Networks that try to predict the winner of a Tennis-Match. To accomplish that, at first a dataset with over 200.000 entries and more than 200 variables was split into four chunks to be the input for one neural Network each. Those neural Networks were made and trained in Tensorflow and their accuracy was tested via picking three players and trying to predict their matches. The results were then plotted into a Scatterplot. With this methodology it is possible to get information about what variables, that the public has access to before the match starts, are how important for the outcome of a match. Furthermore it would be feasible to find out, what set of variables is best suited to make predictions about the winner and also if there are differences in the predictions of the three players.

Contents

Kurzfassung **vii**

Abstract **ix**

Contents **xi**

1 Einleitung **1**
1.1 Fokus Tennissport 1
1.2 Machine Learning 2
1.3 Machine Learning für Vorhersagen 4
1.4 Machine Learning im Sport 6
1.5 Machine Learning im Tennissport 6

2 State Of The Art **9**
2.1 Machine Learning für Vorhersagen 9
2.2 Machine Learning für Ergebnisvorhersagen im Sportsektor 13
2.3 Vorgehensweise der Wettanbieter/innen 17
2.4 Visualisierung 22
2.5 Frameworks und Libraries 26

3 Die Applikation **27**
3.1 Datensatz 27
3.2 Methodik 30
3.3 Visualisierung 37
3.4 Resultate 40

4 Diskussion **49**
4.1 Einblicke und Schlussfolgerungen 49
4.2 Anwendungsgrenzen 53
4.3 Zukünftige Forschung 54

Abbildungen **57**

Tabellen **59**

CHAPTER 1

Einleitung

Diese Arbeit beschäftigt sich mit dem Einsatz von neuronalen Netzen in der Vorhersage von Tennis-Matches auf der höchsten Spielklasse der Herren, der ATP (= association of tennis professionals). Als erstes wird eine Literaturrecherche vorgenommen, die erschließen soll, ob, beziehungsweise inwiefern "Machine Learning" für die Teilbereiche dieser Arbeit genutzt werden kann. Als erstes wird Machine Learning generell definiert und die genutzte Methode erläutert, dann wird die Anwendung von Machine Learning für Vorhersagen, dann die Nutzung selbiger im Sport-Setting, und schließlich der Einsatz im Tennissport thematisiert und ausgearbeitet.

1.1 Fokus Tennissport

Der Fokus dieser Arbeit und der damit zusammenhängenden Webapplikation ist bewusst auf den Tennissport gelegt worden, da sehr viele Faktoren berücksichtigt werden müssen, um Erfolg im Tennissport vorhersagen zu können. In der Vorgehensweise von Panjan et al. [PŠF10] wurde versucht aufgrund von verschiedenen Merkmalen, die in psychische, körperliche und ergebnissorientierte Merkamale eingeteilt wurden, von Jugendlichen mit Hilfe von Machine Learning vorherzusagen, wie erfolgreich diese Jugendlichen später werden würden. Hierbei wurden viele Faktoren wie Größe, Gewicht, Muskelmasse und auch die Leistung in verschiedenen Kraft-, Ausdauer-, und Schnelligkeitstests miteinbezogen. Trotzdem war es ihnen nicht möglich eine Genauigkeit zu erzielen, die zufriedenstellend, geschweige denn besser als die von Scouts oder Trainern, wäre. Die komplexen und vielseitigen Faktoren, die in die tennisbezogene Leistung miteinfließen, werden als Grund genannt. Tennis ist ein Sport, in dem die Leistung von internen Faktoren wie Technik, Taktik, Fitness, Psychologie und Genetik, aber auch von vielen externen Faktoren wie dem Trainingsumfang, den Trainingsmethoden, dem Trainingsumfeld, dem Druck von Eltern und/oder Trainer/in und auch der Anzahl und dem Niveau der gespielten Turniere beeinflusst wird. Die mentalen Faktoren sind besonders im Tennissport sehr

relevant, es geht viel um Konzentration, Selbstregulierung und auch Motivation. Aber auch die motorischen Bewegungen müssen automatisiert und die Intuition geschult werden. Und das muss alles in extremen Drucksituationen und unter höchsten körperlichen Belastungen für den/die Sportler/in abrufbar sein. Ebenfalls müssen für konstante Höchstleistungen die Emotionen der Athlet/innen reguliert und unter Kontrolle gehalten werden, beispielsweise Ärger, der in einer Spielsituation ungewollte Anspannung erzeugen kann, was den/die Spieler/in und seine/ihre Leistung hemmt [Grä03]. Generell liegt eine Korrelation zwischen mentaler Stärke und der Leistung im professionellen Bereich des Tennissports, abhängig von der spezifischen Matchsituation, sehr nahe [Cow16]. Darüber hinaus suggeriert die Arbeit von Cowden et al. [CMWOA16] eine stresslindernde Wirkung von höherer mentaler und psychischer Stärke, was in der Matchsituation noch zusätzlich die Leistung in Drucksituationen steigern kann.

Diese Komplexität ist es auch, was es so interessant macht, Erfolgsprognosen im Tennis mit Machine Learning durchzuführen. Die Frage, die es zu beantworten gilt, ist, ob die vielen Variablen in einem Zusammenhang stehen oder nicht, und ob damit eine Vorhersage wirklich auch möglich ist [PŠF10]. Da verschiedenste mentale Instabilitäten oder fehlende Selbstregulierungsmaßnahmen zu starken Leistungsschwankungen führen können und auch die emotionale Tagesverfassung und viele andere Faktoren, auf das Ergebnis eines Matches einwirken [Grä03], ist es ebenfalls denkbar, dass eine objektive Vorhersage aus den vorherigen Matchdaten nicht möglich ist, da zu viele wichtige Faktoren nicht in den Daten vorhanden sind. Durch die Visualisierungsapplikation, die in dieser Arbeit implementiert wird, sollen Faktoren herausgefiltert werden, die für die Bestimmung des Ausgangs eines Tennis Spiels am relevantesten sind. Damit sollen sich bestenfalls Trends oder Zusammenhänge herausfiltern lassen, die für Wettanbieter/innen und die Quotenerstellung nützlich sind. Aber auch Trainer/innen und Spieler/innen sollen durch konkretere Trainingsziele aufgrund von klareren Variablen, die ein Match entscheiden, profitieren. Bis aufgrund von zukünftigen Arbeiten, die Objektivität von Machine Learning ideale Bedingungen am Wettmarkt und auch im Trainingsalltag bieten kann [ROD21].

1.2 Machine Learning

Machine Learning ist eine Sparte von Algorithmen, die die Intelligenz und Denkfähigkeit der Menschen imitiert. Dies geschieht, indem Algorithmen von ihrer Umgebung (Trainingsdaten) lernen und damit Klassifizierungen und Entscheidungen treffen. Jedoch hat diese Vorgehensweise bereits in einer beträchtlichen Anzahl von Branchen zu Erfolgen geführt, sei es in Computer Vision, Mustererkennung, Raumfahrt, am Finanzmarkt oder auch in der Radiotherapie [ENM15]. Mittlerweile haben sich zahlreiche verschiedene Arten der Machine Learning Algorithmen gebildet, die in manchen Einsatzgebieten bessere oder schlechtere Leistung liefern. Dies umfasst unter anderem Lineare Regression, Genetische Algorithmen, "support vector machines" (= SVM), Fallbezogene Argumentation (= "case-based reasoning") und, die, in dieser Arbeit genutzten, Neuronalen Netze [YKJ05]. Alle diese Algorithmen automatisieren und optimieren diverse Prozesse, von

der Qualitätssicherheit bis hin zur Planung und werden auch, wie in dieser Arbeit von besonderem Interesse, zur Vorhersage eingesetzt [ENM15]. Dennoch gibt es Gebiete, die aufgrund ihrer Eigenheiten noch mehr Forschung in die Richtung von Machine Learning benötigen, um es zu seinem vollen Potential und wirklich effizient nutzen zu können, beispielsweise in der Aufwandseinschätzung von Software-Projekten [MKL+00].

1.2.1 Neuronale Netze

In dieser Arbeit wird besonderer Fokus auf den Machine Learning Ansatz der neuronalen Netze gelegt, da auch im praktischen Teil mit solchen neuronalen Netzen gearbeitet wird. Neuronale Netze simulieren die Funktion des menschlichen Hirns, indem sie die Neuronen und Synapsen nachbilden [ENM15]. Die Neuronen sind in diesem Fall einzelne Aktivierungsfunktionen, die entweder feuern, oder nicht feuern. Diese Neuronen werden nun, wie in Abbildung 1.1 zu sehen ist, in mehreren Schichten (layers) angeordnet. Diese Schichten erfüllen unterschiedliche Funktionen, die Eingabeschicht (= input layer) nimmt die Eingangsparameter des Netzwerks auf und gibt sie, über die Synapsen (in diesem Kontext Gewichte oder weights genannt) an die Neuronen der Versteckten Schicht weiter. Hierbei ist normalerweise jedes Neuron dieser beiden Schichten miteinander verbunden und die Werte der Synapsen entscheiden, wie viel Einfluss das jeweilige Neuron der vorherigen Schicht hat. Von diesen versteckten Schichten, kann es, je nach Komplexität des Problems, mehrere oder auch nur eine geben. Jedenfalls werden die Ergebnisse der letzten versteckten Schicht an die Ausgabeschicht weitergegeben. In dieser Schicht repräsentiert nun jedes Neuron einen möglichen Wert, der ausgegeben werden kann. Das ist entweder einer von vielen Klassifikationswerten oder auch, wie in dieser Arbeit von Relevanz, ein Ergebnis einer Vorhersage. Ebenfalls gibt es sogenannte Bias Neuronen, die einen konstanten Wert an die Neuronen weitergeben, der über das Training angepasst werden kann. Training bedeutet, dass das neuronale Netz Daten aus einem Trainingsdatensatz als Eingabe bekommt, deren Klassifikation oder Vorhersage (ground-truth = GT) bereits bekannt ist, dann wird mit der GT verglichen, wie viele Werte richtig klassifiziert oder vorhergesagt wurden. Im nächsten Schritt werden dann die Konstante des Bias und auch die Werte der Gewichte angepasst, um bessere Ergebnisse erzielen zu können und es wird wieder versucht, anhand der Trainingsdaten eine gute Leistung zu erzielen. Dieser Vorgang und das Anpassen wird eine fixierte Anzahl an Durchgängen (=Epochen) durchgeführt, bis das System eine akzeptable Quote an richtig klassifizierten/vorhergesagten Daten vorweisen kann [AJO+18].

Aufgrund dieser Eigenschaft können einem neuronalen Netz ganz unterschiedliche Aufgaben gegeben werden, solange es jeweils nur diese eine Aufgabe durchführen muss. Beispielsweise werden neuronale Netze in Spracherkennung, Musterekennung, Datenrekonstruktion, Gesichts- und Schrifterkennung [AJO+18], aber auch zur Vorhersage von beispielsweise Aktienkursen auf der Börse [YKJ05] eingesetzt. Wobei meist die Fähigkeit komplexe, nichtlineare Zusammenhänge erkennen zu können eine große Rolle spielt. Über die Jahre wurden auch noch spezielle neuronale Netze konzipiert und sind bereits im Einsatz. Beispiele hierfür sind die "Feed Forward" neuronalen Netze, die bereits in

Basketball-Vorhersagen erfolgreich waren [LBB09]. Zu dieser Gruppe zählen das Multi- und Single Schichten Perceptron und auch radiale Basisfunktionsnetzwerk. In Feed Forward neuronalen Netzen sind alle Neuronen einer Schicht, mit allen Neuronen der nächsten Schicht verbunden. Wogegen die zweite Art, die Feed Backward neuronalen Netze, den internen Status nutzt, um eine Sequenz an Eingaben zu verarabeiten. Eine Subvariante davon sind die wiederkehrenden neuronalen Netze (recurrent neural Networks), die bereits im Fußball bei der Ergebnisvorhersage eingesetzt wurden [JTS21]. Weitere Varianten sind das Deep Learning, das prinzipiell neuronale Netze mit komplexeren und zahlreichen versteckten Schichten beschreibt, dann noch gefaltete (convolutional), auch im Einsatz im Bereich des Fußballs [HŠŽ19], und auch rekursive neuronale Netze.

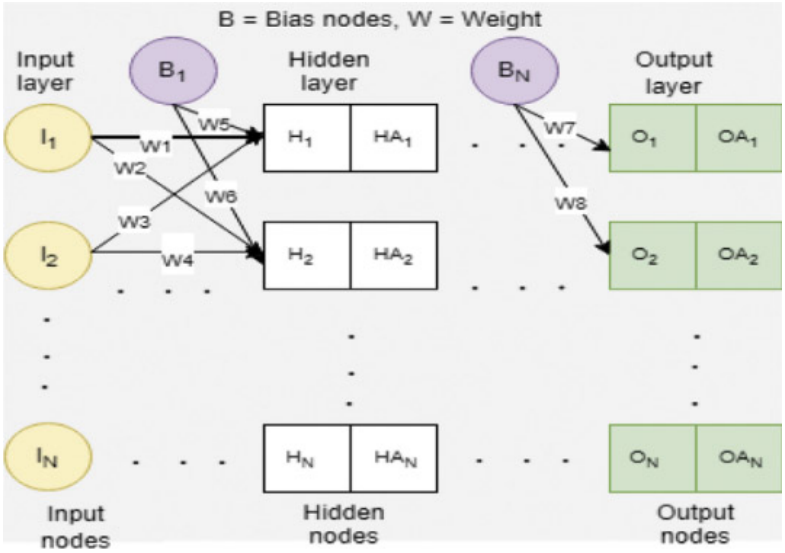

Abbildung 1.1: Typische Architektur eines neuronalen Netzes. Bildquelle: [AJO+18]

.

1.3 Machine Learning für Vorhersagen

Für Vorhersagen werden Machine Learning Algorithmen häufig eingesetzt. Machine Learning kann entweder für Klassifikationen genutzt werden, hierbei wird beispielsweise ein Bild analysiert und es wird bestimmt, um was für ein Bild es sich handelt, im Vorhinein muss der Algorithmus mit ähnlichen Trainingsdaten trainiert werden, damit diese Klassifikation korrekt ist [AJO+18]. In der Vorhersage ist es ähnlich, hierbei werden jedoch zuerst Werte aus der Vergangenheit genommen um zu trainieren, bei denen das Ergebnis schon feststeht. Und dieses Training wird dann auf neue Situationen angewendet, die noch nicht passiert sind. Somit wird das Ergebnis dieser kommenden Situation vorhergesagt und es wird erst im Nachhinein klar, ob die Prognose richtig war. Beispielsweise in der Krebsforschung, wo die richtige Klassifizierung von Patient/innen in "Patient/in mit geringem Risiko" und "Patient/in mit hohem Risiko" essenziell ist,

werden verschiedenste Arten des Machine Learnings eingesetzt, um den Verlauf von krebsartigen Zuständen und ihre Behandlung vorherzusagen. Hierbei spielt die Fähigkeit des Machine Learnings, Schlüsseleigenschaften aus großen und komplexen Datensätzen herauszufiltern, eine große Rolle [KEE+15]. Wie die Ergebnisse von Manikandan und Chinnadurai [MC20] zeigen, kann mit Hilfe des Machine Learning Ansatzes Deep Learning auch eine 85 - 90 prozentige Genauigkeit erreicht werden, wenn man die zukünftigen Leistungen von Studenten (sowohl akademisch als auch sportlich) vorhersagen will. Hierfür werden verschiedene Fragebögen ausgefüllt und ausgewertet, wobei die Genauigkeit der Vorhersagen für Interessensgruppen ausreicht und Manikandan und Chinnadurai zu dem Schluss kommen, dass diese Art der Vorhersage mit Hilfe von Machine Learning zukünftig mehr Einsatz in breitgefächerten Bereichen der Ausbildung und auch im Berufsleben finden wird.

Ebenfalls eingesetzt werden diese Algorithmen laut Yoo et al. [YKJ05] im Finanzsektor um Aktienkurse vorhersagen zu können. In diesem Einsatzgebiet haben sich nach längerem Testen die neuronalen Netze als Machine Learning Algorithmen mit der besten Leistung herauskristallisiert. Und es hat sich auch bewährt, dass neuronale Netze bessere Ergebnisse zeigen, als herkömmliche statistische Methoden. In diesem Kontext ist die Fähigkeit von neuronalen Netzen, nicht lineare Zusammenhänge zwischen Trainingsdatensätzen und deren Labels (= richtigen Klassifizierungen) zu erkennen, von hoher Bedeutung, da der Aktienmarkt ein komplexes, dynamisches und nicht lineares System darstellt. Durch diese Komplexität ist ein nicht lineares Modell von Nöten, um das System so gut es geht abbilden zu können. Darüber hinaus benötigt es meistens eine große Anzahl an miteinander interagierenden Input-Variablen, um Aktien darzustellen. Und während andere Varianten, wie traditionelle statistische Modelle, eine funktionale Form der Beziehungen von Variablen annehmen müssen, können neuronale Netze diese aus den Trainingsdaten selbst herausfinden. Als Nachteil der neuronalen Netze in dieser Einsatzmöglichkeit nennen Yoo et al. [YKJ05] beispielsweise das "black-box"-Problem dieser Vorgehensweise. Das bedeutet, dass nicht ersichtlich ist, wie und wieso einzelne Variablen so gewichtet werden, was es unmöglich macht, die Entscheidungen - in diesem Kontext die Berechnungen der zukünftigen Aktien-Preise - des neuronalen Netzes nachzuvollziehen.

Auch in anderen Bereichen der Wirtschaft ist Machine Learning sehr gefragt. Beispielsweise werden diverse Ansätze bereits in der Prognose von Marktpreisen eingesetzt, damit sich ein Unternehmen am potentiellen Preis der Mitbewerber und auch dem Kundenverhalten orientieren kann [MIS+21]. Außerdem wird Machine Learning und insbesonders neuronale Netze aufgrund ihrer Eignung für Mustererkennung [AJO+18] für sogenanntes "Data-Mining" eingesetzt. Das ist der Prozess, bei dem aus großen Datenmengen, wie die Nutzungsstatistiken von Plattformen wie Google oder Facebook Informationen herausgefiltert werden sollen, die relevant sind für personifizierte Werbung oder Ähnliches [BM01].

1.4 Machine Learning im Sport

Machine Learning hat im Sportsetting zahllose Einsatzmöglichkeiten und in einigen Bereichen ist Machine Learning bereits sehr einflussreich. Beispielsweise, wie Richter et al. [ROD21] beschreiben, die Datenverarbeitung, hier wird Machine Learning bereits in den Geräten eingesetzt, die die Daten erheben, um das Lesen der Sensor-Rohdaten zu optimieren. Ebenfalls betonen die Autoren, dass es eingesetzt wird, um relevante Informationen aus den Daten herauszufiltern, diese dann im nächsten Schritt aufzubereiten, was zuvor viel Zeitaufwand für die zuständigen Sportwissenschaftler/innen bedeutete, und darüber hinaus hilft Machine Learning, diese Daten dann zu nutzen, um Leistungen und Verletzungen vorhersagen zu können. Wie die Resultate der Studie von Rommers et al. [RRV+20] zeigen, ist Machine Learning in der Lage, mit 85 Prozent Präzision (ebenfalls 85 Prozent Genauigkeit und auch Sensitivität) Verletzungen korrekt vorherzusagen und die Verletzungen mit 78 Prozent Präzision, Genauigkeit und Sensitivität auch richtig nach Verletzungsursache in "Überbeanspruchung" oder "Akute Verletzung" zu klassifizieren. Dies schafft der Algorithmus basierend auf den Messungen, die in der Saisonvorbereitung gemacht wurden. Diese Ergebnisse legen nahe, dass dieses Einsatzgebiet in Zukunft noch weiter ausgebaut und dadurch die Objektivität in sportwissenschaftlichen Entscheidungen stark steigen wird. Trotz der Verbesserungschancen müssen auch gewisse Herausforderungen überwunden werden, damit diese Entwicklung geschehen kann. Beispielsweise können verlorene Informationen (beispielsweise die Zeiten zwischen den Blutabnahmen bei einem Laktatstufentest) ein Problem für die Vorhersage und die Genauigkeit des Algorithmus bedeuten, aber auch Over- und Underfitting - also das zu starke Anpassen an die Trainingsdaten (over-) und das zu allgemeine, also zu schwach auf die Trainingsdaten bezogene, Training (under-) - sind Probleme, die bedacht werden müssen [ROD21].

Ebenfalls ein großer potentieller Einsatzbereich von Machine Learning im Sport ist der Wettmarkt, wobei die meisten herkömmlichen Wettanbieter/innen immer noch ihre Quoten von Buchmacher/innen, also Expert/innen, beziehen, gibt es bereits Ergebnisse, die schon einen leichten und steigenden Vorteil zu Gunsten der Machine Learning Ansätze impliziert [FHS18]. Jedoch ist es unwahrscheinlich, dass eine einheitliche Lösung für alle Sportarten gefunden wird, da die massiven Unterschiede zwischen den Sportarten zu sehr unterschiedlichen Ergebnissen in der Vorhersage von Machine Learning Ansätzen führen [HJ20].

1.5 Machine Learning im Tennissport

Auch im Tennissport wurden bereits Versuche durchgeführt, mit Machine Learning in verschiedenen Bereichen des Sports eine höhere Objektivität und Automatisierung zu erreichen. Beispielsweise haben Panjan et al. [PŠF10] versucht mit Hilfe von Deep Learning und anderen Machine Learning Ansätzen vorherzusagen, wie erfolgreich die 12- bis 16-jährigen Tennisspieler/innen aus Slowenien in Zukunft sein werden, jedoch waren diese Vorhersagen sehr ungenau und die Methode daher nicht von Erfolg gekrönt. Hilfreich war Machine Learning hingegen beispielsweise bei der Erkennung und Klassifizierung von

Assen (Aufschläge, die nicht mehr vom Rückschläger berührt wurden). Hierfür wurde eine große Anzahl von Aufschlägen während des Australian Open Turniers zwischen 2012 und 2015 analysiert, ausgewertet, und die wichtigsten Kriterien herausgefiltert. Dies ist deshalb von Nutzen, da dieses Wissen Trainer/innen zur Verfügung gestellt werden kann und somit an den wichtigsten Teilbereichen des Aufschlags gearbeitet werden kann, um sowohl die Leistung, als auch die Trainingseffizienz zu steigern. Der Aufschlag in höheren Klassen muss immer mit genügend Platzierung und auch trotzdem Schnelligkeit in der Bewegung und folglich auch Tempo und Drall im abgeschlagenen Ball ausgestattet sein. Gerade beim Aufschlag können Millimeter entscheiden, weshalb es für Trainer/innen und ihre Schützlinge sehr von Nutzen ist, das Training anhand von berechneten Idealwerten, wie in Abbildung 1.2 zu sehen, gestalten zu können, um die maximale Leistung aus diesem wichtigen Schlag herauszuholen [WR17].

Zusätzlich wurden auch Studien durchgeführt, die mit Hilfe von Machine Learning verschiedene Bewegungen von Tennisspieler/innen analysierten. In der Arbeit von Giles et al. [GPKR21] wurde beispielsweise ein spezieller Fokus auf die Richtungswechsel ("change of directions") gelegt, während Liu X. [Liu20] mit Hilfe von Sensorik und von "Support Vector Machines", einem Machine Learning Modell, sich die Klassifizierung der verschiedenen Schläge im Tennis (Vorhand, Rückhand, Aufschlag, etc.) zum Ziel gemacht hatte, welches auch rund 79 Prozent Genauigkeit erreichen konnte.

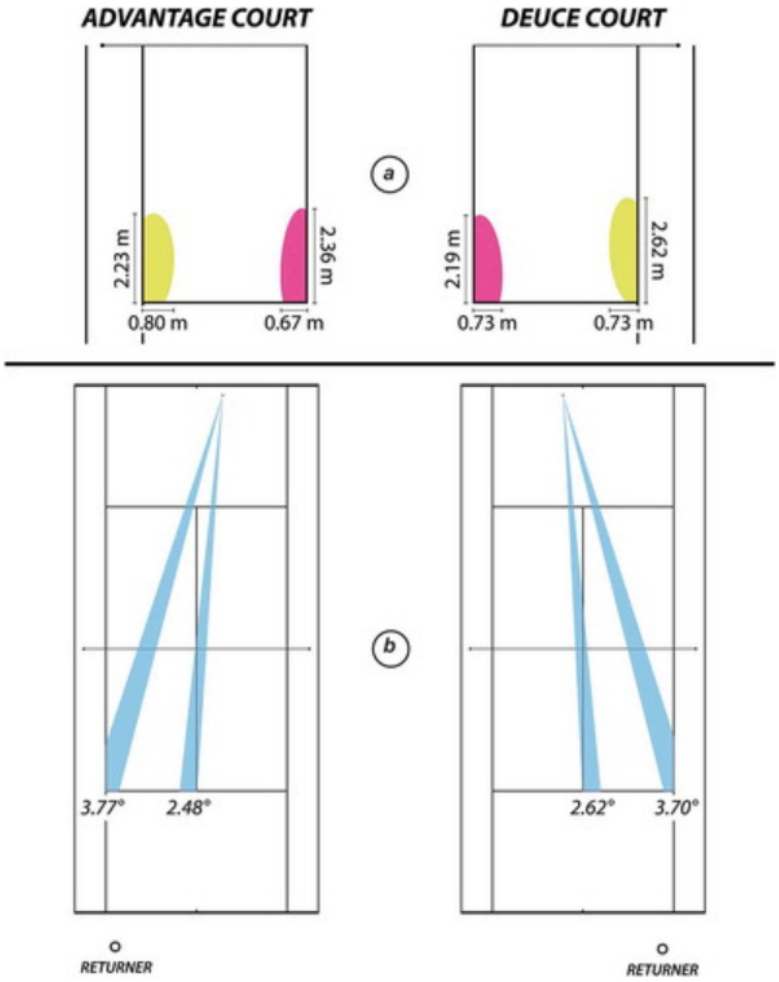

Abbildung 1.2: Machine Learning für die Berechnung der optimalen Aufschlagplatzierung im Tennis Einzel. Bildquelle: [WR17]

.

CHAPTER 2

State Of The Art

Teil dieser Arbeit ist es, einen Überblick zu geben, wo und wie Machine Learning bereits im Einsatz ist, um eine Basis für den praktischen Teil zu liefern. Hierfür werden die einzelnen Teilgebiete des Themas einzeln betrachtet und Vorkommnisse von Machine Learning in der Literatur dokumentiert. Darüber hinaus wird darauf eingegangen, wie Wettanbieter/innen ihre Quoten bestimmen, um zu analysieren, wie Machine Learning, und konkreter die vorgestellte Applikation, im Vergleich dazu abschneiden. Machine Learning ist in den letzten 20 Jahren immer erfolgreicher geworden und viele Sparten setzen bereits Machine Learning Verfahren, wie Neuronale Netze, Regression, Datenbündelung, ein, Tendenz schon 2001 steigend [MD01]. Zusätzlich wird auf gängige Visualisierungmethoden in diesem Kontext, relevante Frameworks und Libraries für Machine Learning und auf auch zukünftige Herausforderungen eingegangen.

2.1 Machine Learning für Vorhersagen

2.1.1 Biologie

Bereits 1997 wurden Erfolge mit Hilfe von Machine Learning im Bereich der Genetik verzeichnet. Burge und Karlin [BK97] haben damals ein Programm vorgestellt, das mehrere Gene in einer Sequenz und auch konsistente Gensets der DNA Stränge vorhersagen kann. Um das zu schaffen, nimmt das probabilistische Modell, das in ihrem Programm zum Einsatz kommt verschiedenste Daten der Gene als Parameter, wie beispielsweise verschiedene Signale, Längenverhältnisse und auch Eigenschaften des Aufbaus von intergenischen Regionen, Exonen und auch Intronen. Diese Parameter werden dann abgeleitet, um die Unterschiede der Regionen des menschlichen Genoms auszugleichen und dann wird die Vorhersage durchgeführt. Die Genauigkeit des Programms übertraf in jedem relevanten Teilbereich die, der bisher zur Verfügung stehenden Applikationen. Darüber hinaus argumentieren Burge und Karlin, dass sich das Einsatzgebiet für

eine effektive Methode, Gene aus einer anonymen DNA heraus identifizieren zu können, stark vergrößern wird. Dafür ist Machine Learning ideal, da durch diese Vorgehensweise ein großer Teil der Suche nach den Exon-Aufenthaltsorten eingespart werden kann. Dieser Einfluss von Machine Learning auf die Genetik und Genomik hat sich in den letzten Jahren nur verstärkt und die Einsatzgebiete wurden ausgeweitet. Es werden nun bereits Funktionen und Schädlichkeit von Genen anhand der DNA Sequenz und den Histon-Modifikationen vorhergesagt. Darüber hinaus werden Krankheits-Phenotypen klassifiziert und somit potentielle biologische Marker von Krankheiten im Vorhinein identifiziert und auch Beziehungen zwischen Genen können vorhergesagt werden [LN15].

2.1.2 Medizin

Machine Learning ist ein essentieller Bestandteil für die Arbeit mit medizinischen Bildern. Sowohl in ihrer Analyse, als auch in der Computer unterstützten Disagnose (CAD). Das kommt daher, dass sowohl Läsionen als auch Organe in medizinischen Bildern zu komplex erscheinen, als dass sie von einfachen statistischen Methoden klassifiziert werden könnten. Konkret werden sogenannte Pixel basierte Machine Learning Algorithmen (bedeutet, jedes Pixel und seine Eigenschaften sind ein Eingangsparameter für diesen Algorithmus) dazu verwendet, um gewisses Quantenrauschen aus Bildern zu entfernen, die durch Diagnose-Röntgen aufgenommen wurden. Dafür sind besonders neuronale Netze geeignet, da sie die Bilddetails, wie beispielsweise Kanten, gut erhalten können. Ebenfalls können diese Algorithmen als Kantenverbesserer für Ausgangsbilder, die ein derartig hohes Rauschen aufweisen, dass die Kanten so nicht weiterverarbeitbar wären, eingesetzt werden. Darüber hinaus können auch verschiedene Teile des menschlichen Körpers auf Röntgenbildern erkannt und separiert werden. Beispielsweise können Kochen von Bindegewebe in einem Brust-Röntgenbild getrennt werden, um so leichter Krankheiten im Brustbereich wie Lungenkrebs besser und früher zu diagnostizieren, da in der Vergangenheit sehr viele Fälle von Lungenkrebs übersehen wurden, da sie von Knochen verdeckt wurden. Das Potential von Machine Learning in dem Bereich der Diagnose ist noch viel größer, da es möglich ist, Läsionen besser zu erkennen und auch verschiedene Arten zu klassifizieren. Auch in diesem Bereich werden hauptsächlich verschiedene Arten von neuronalen Netzen eingesetzt, vor allem gefaltete und 'shift-invariant' neuronale Netze. Letztere sind neuronale Netze, bei denen ein Vertauschen der Eingabedaten zu vergleichbaren Änderungen in der Ausgabe führt [Suz12].

Im Fachbereich der Kardiologie wurde Machine Learning und künstliche Intelligenz immer präsenter, vor allem in der Forschung. Wie Quer et al. [QAHA21] herausfanden, ging die Zahl der Publikationen mit dieser Themenkombination (in den behandelten Magazinen) stark nach oben in den letzten Jahren. Sehr viele der publizierten Artikel drehen sich um Atherosklerose, Hypertension, Herzversagen und andere kardiologische Risikofaktoren und die meisten arbeiten mit neuronalen Netzen. Nur vereinzelt sieht man auch herkömmliche statistische Methoden, wie beispielsweise Entscheidungsbäume oder lineare Regression. Machine Learning wird hier für diverse Vorhersagen eingesetzt, beispielsweise Risiko-Vorhersagen von myokardialen Infarkten. Ebenfalls können dann

klinische Daten genutzt werden, um genetische Varianten, systolischen, bzw. diastolischen Blutdruck und sogar das Sturzrisiko - durch tragbare Sensoren - vorherzusagen. Hinzu kommt, dass Machine Learning auf einer Systemebene im Gesundheitssystem eingesetzt wird, indem beispielsweise eine Untersuchung auf kardiovaskuläre Krankheiten durch Algorithmen ausgewertet wird. Mit der Hilfe dieser Daten kann dann ein optimaler Behandlungsplan erstellt werden und auch Fehlalarme reduziert werden.

Auch im Hinblick auf Covid-19 wird Machine Learning eingesetzt. 2021 haben Bottino et al. [BTP+21] beschrieben, dass vor allem Machine Learning-Ansätze, die es schaffen, nichtlineare komplexe Interaktionseffekte zu modellieren, die Vorhersage der Tödlichkeit von Covid-19 stark verbessern. Ebenfalls findet Machine Learning Verwendung, um die Anzahl neuer Infektionen vorherzusagen, indem die Verbreitung des Virus modelliert wurde. Mit sehr vielen Ansätzen wurde dafür experimentiert, jedoch stellte sich ein gemischtes Netzwerk (bestehend aus nichtlinearer Autoregression und funktionsanpassenden neuronalen Netzen) und ein "long short-term memory" (= LSTM) neuronales Netz als beste Herangehensweisen heraus. Letzteres zeichnet sich durch das Einbeziehen von den Ergebnissen aus mehreren vorherigen Schichten in die Entscheidungen der aktuellen Schicht aus. Zeichen einer Covid-19 Infektion können auch von Machine Learning Methoden auf dem Lungenröntgenbild des/der Patient/in erkannt und diagnostiziert werden, wobei Deep Learning mit Hilfe dieses Bruströntgens eine Genauigkeit von 99 Prozent erreicht. Und sogar der Bereich der Pharmazie kann durch die gezielte Modifikation von existierenden Grippeimpfstoffen mit Machine Learning schneller zu einer potentiellen Covid-19 Impfung kommen. Hierfür wurden mehrere verschiedene Ansätze mit mehreren verschiedenen Trainingsdaten analysiert und dann der Ansatz mit der besten Leistung herangezogen. Dann wurde mit diesem getestet, wie die Zielpersonen auf die jeweiligen Medikamente reagieren und ob die Reaktion im Kontext einer möglichen Covid-19 Infektion relevant sein könnte. So wurden die, für die Basis in Frage kommenden, Medikamente aussortiert [KCS+21].

2.1.3 Wirtschaft

Bereits in der Agrarwirtschaft gibt es zahlreiche Teilbereiche, die von Machine Learning Algorithmen profitieren. Durch die globale Erwärmung und unvorhersehbare Regenmuster sind Bauern/Bäurinnen mit herkömmlichen Methoden zum Anbau oft dem Zufall ausgeliefert. Die Hilfe von Machine Learning hingegen kann sogenannte Präzisions-Landwirtschaft ermöglichen, die den Farmer/innen in jedem Schritt des Anbauprozesses unterstützt. Schon im ersten Schritt, der Auswahl, welche Pflanzen, wann, wo und in welcher Menge angebaut werden können, erleichtert eine Vorhersage eines Algorithmus die Arbeit. Indem die geographischen und klimatischen Beschaffenheiten der Landschaft als Parameter eingegeben werden, können die Bodeneigenschaften vorhergesagt werden und somit die Bepflanzung präzise geplant werden [SJGC21]. Sind die Pflanzen dann gesät, so hilft Machine Learning, um Krankheiten zu erkennen, damit frühzeitig Gegenmaßnahmen, wie Pestizide, ergriffen werden können. Auch bei den Pestiziden unterstützt auch wieder das Machine Learning, da es vorhersagt, welches Pestizid nun eingesetzt werden

sollte, um den idealen Effekt zu erzielen, dafür werden aktuell Support Vector Machines und Extreme Learning Machines eingesetzt. Im Ernteprozess hilft dann schließlich ein mit Kamera und Sensoren ausgestatteter Roboter mit Deep Learning um Verluste zu reduzieren [MPM+21].

Im Verkauf wird dann häufig Machine Learning eingesetzt, um Preisvorhersagen von relevanten Mitbewerbern zu erstellen und somit die eigenen Preise daran anpassen zu können. Zusätzlich kann auch aufgrund von Konsument/innendaten der Bedarf an gewissen Produkten vorhergesagt werden. Das vollständige Modell nimmt dann alle diese Daten und zusätzlich die Kaufhistorie der Kunden, das Feedback und die Transaktionen als Parameter und es werden Vorschläge gemacht, die die aktuellen Preise des Unternehmens so verändern, dass ein Vorteil für das Unternehmen entsteht. Beispielsweise kann ein Rabatt auf ein Produkt mit nicht so hohem aktuellen Bedarf vorgeschlagen werden, um einen Kaufreiz zu setzen. Durch diese Entscheidungen kann sowohl die Produktivität als auch der Marktanteil gesteigert werden. In diesem Einsatzbereich war die Einer-vs-Alle Multiklassen Methode (all-vs-one multiclass method) am erfolgreichsten [MIS+21]. Generell wurde die gesamte Planung eines Unternehmens durch Machine Learning unterstützte Business Intelligence Software verändert. Es werden Markteigenschaften, Mitbewerberverhältnisse und Kund/inneneigenschaften vorhergesagt und analysiert und aufbauend auf diesen Daten werden dann Entscheidungen getroffen [TROKRP21].

Auch Banken profitieren von Machine Learing, da sie durch die künstliche Intelligenz größere Mengen und verschiedene Datentypen von Variablen in ihre Kreditwürdigkeitsanalysen miteinbeziehen können und somit die Genauigkeit dieser Analysen steigt. Banken sind darauf aufmerksam geworden, dass verschiedenste, unwichtig erscheinende Variablen wie Telekom Daten oder Daten von sozialen Medien diese Analysen deutlich verbessern. Diese Anpassungen sind laut Königstorfer und Thalmann [KT20] wichtig für die Zukunft der Banken, da sie die große Anhäufung von Daten, die für die Klassifizierung durch Machine Learning notwendig ist, bereits alltäglich durchführen und so einen Vorteil gegenüber Nicht-Banken als Konkurrenz haben. Darüber hinaus werden die Algorithmen durch die neuen Variablen genauer und so können die Banken eine größere Anzahl von Kreditnehmer/innen aufnehmen, was nochmals Profit bringt.

2.1.4 Probleme und Herausforderungen

Eines der größten Probleme des Machine Learnings ist die Notwendigkeit, dass Menschen händisch die Vorarbeit für das Modell leisten müssen, es müssen die Daten zusammengetragen, ausgewertet und richtig gefiltert sein. Jedoch hört es da nicht auf, auch die Hyperparameter des Modell müssen angepasst werden, sonst können keine idealen Ergebnisse erzielt werden. Diese Anpassung geht auch nur händisch und benötigt meist sehr viele Versuche, um eine ideale Hyperparameterwahl zu treffen. In Zukunft wird an einer Variante geforscht werden, die erlaubt, dass zumindest das Hyperparameter-Anpassen wegfällt und dass künstliche Intelligenz diesen Schritt durchführt, ohne menschliche Hilfe. Das ist einer der relevantesten Punkte für die Zukunft dieser Forschung und

Ergebnisse könnten die Arbeit mit Machine Learning Algorithmen deutlich vereinfachen und verbessern [EMS19].

2.2 Machine Learning für Ergebnisvorhersagen im Sportsektor

Aufgrund der Unterschiede in unter anderem Belastungen, körperlich und mental, Teamdynamiken, Wettbewerbsdurchführung und mehr, sind verschiedene Machine-Learning-Verfahren in verschiedenen Sportarten verschieden erfolgreich. Bei manchen Sportarten gibt es weniger Diskrepanzen bei den Genauigkeiten der verschiedenen Herangehensweisen, bei anderen sind beinahe 40 Prozent Schwankungen möglich, wie man in Abbildung 2.1 betrachten kann [HJ20].

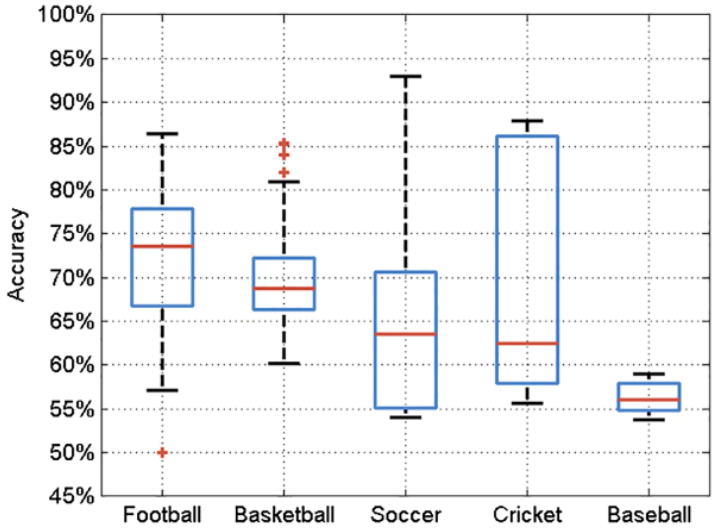

Abbildung 2.1: Unterschiede in den Genauigkeiten von Machine Learning in verschiedenen Sporarten via Kastengrafik. Bildquelle: [HJ20]

2.2.1 Basketball

Bereits 2009 wurde von Löffelholz et al. [LBB09] ein Konzept entwickelt, das die Ergebnisse von Basketball-Matches in der NBA (National Basketball Association) vorhersagen kann. Hierfür wurden auch wieder mehrere Machine-Learning-Technologien getestet, die beste Leistung brachte in dem Fall eine Fusion aus vier verschieden aufgebauten, neuronalen Netzen, die in ein Bayesisches Modell kombiniert werden und dann gemeinsam das Siegerteam ermitteln. Input für die neuronalen Netze waren die Ergebnisse und Statistiken der Matches aus der Saison 2007-2008. Diese Forschung zeigte, dass besonders Feedforward-Neuronale-Netze, probabilistische neuronale Netze und auch generalisierte

neuronale Netze sich gut für diese Vorhersagen zu eignen scheinen. Durchschnittlich erreicht das vorgestellte System 74,33 Prozent Genauigkeit und im optimalen Fall sogar 83,33 Prozent. Diese Ergebnisse wurden dann mit den Vorhersagen von Expert/innen verglichen und hier war der Algorithmus erfolgreicher. Damals waren die Algorithmen für diese Art der Vorhersage noch nicht weit genug entwickelt [Cao12], bis 2019 wurden diese Applikationen dann weiterentwickelt, und Thabtah et al. [TZA19] kamen bereits auf eine Genauigkeit von 80-83 Prozent, in dem sie durch das Vorbearbeiten der Daten und dem Extrahieren von wichtigen Parametern ihre Ergebnisse verbessert haben.

2021 gelang es dann, mit Hilfe eines Multi-Schichten Perceptron Algorithmus eine Erfolgsquote von über 90 Prozent zu erzielen. Es wurden ebenfalls mehrere Machine Learning Konzepte versucht, jedoch waren das Mulit-Schichten Perceptron und die Logistische Regression den anderen immer deutlich überlegen, mutmaßlich deshalb, da die Datensätze nur aus numerischen Daten bestanden und die Entscheidung immer binär erfolgt. Jedoch sind diese Erfolge mit der Limitierung zu betrachten, das keine amerikanische Liga als Datensatz gewählt wurde, sondern eine europäische. Mit fünf verschiedenen Datenmengen, verschiedener Prozentzahl von verwendeten Daten pro Menge und unterschiedlichen Machine-Learning-Ansätzen und auch Feature-Filterungs-Verfahren wurden Modelle erstellt und ihre Leistung verglichen, um das beste Modell zu finden. Diese Herangehensweise und auch die guten und reichhaltigen Trainingsdaten führten zu diesen hohen Genauigkeitswerten [BÖ21].

2.2.2 Fußball

Im Fußball wird Machine Learning bereits in vielen Bereichen eingesetzt. Die Leistung von einzelnen Spieler/innen kann vorhergesagt werden, sogar für einzelne Spiele. Beispielsweise können anhand der Leistungen der letzten Saison die wahrscheinlichen Leistungen eines/einer Spielers/Spielerin der nächsten Saison mit guter Genauigkeit determiniert werden. Verletzungsrisiken können ebenfalls im Vorhinein identifiziert und somit, so gut es geht, vermieden werden. Außerdem können bereits Talente vorhergesagt werden [RGPOM$^+$22]. Ebenfalls kann Machine Learning für das Vorhersagen von Spielergebnissen oder auch nur Gewinnner/Verlierer/Unentschieden verwendet werden. Da Fußball eine der beliebtesten und finanziell interessantesten Sportarten ist, ist ebenfalls dem Fußball-Wett-Sektor ein großer Wert zugeschrieben. Darum wurde auch bereits viel in der Sparte der Ergebnisvorhersage geforscht und viele verschiedene Machine-Learning-Methoden wurden eingesetzt, um die ideale Leistung zu erhalten. Auch hier wurde eine Feature-Filterung durchgeführt, die die wichtigsten Teile des Datensatzes von irrelevanten trennt. Eine interessante Erkenntnis ist die Relevanz des ersten Tores, das durchschnittlich deutlich bessere Siegeschancen ergibt. Wichtige Parameter für diese Studie waren die Anzahl der Schüsse aufs Tor und auch der Ballbesitz in Prozent. Auf diese Werte wurde folglich auch der Fokus gesetzt. Andere Werte, wie Eckbälle, Fouls und Ähnliches wurde in den Vorbereitungsschritten aussortiert [BU19]. Bereits 2008 wurde von Van Calster et al. [VCSVH08] ein Modell entwickelt, das torlose Unentschieden vorhersagen können soll. Es wurden hierbei sowohl Bayes Multi-Layer Perceptron neu-

ronale Netze getestet, als auch Least-Squares Support-Vector-Maschines (LS-SVM) und es wurden Daten aus verschiedenen Ländern und Ligen im professionellen Fußball verwendet. Wobei das Modell der LS-SVM die bessere Leistung ablieferte. Das Ziel war es herauszufinden, was diese torlosen Unentschieden hervorruft und welche Eigenschaften bei den Mannschaften gegeben sein müssen, um die Wahrscheinlichkeit dieses Ergebnisses zu beeinflussen. Diesen Ansatz wählten auch Geurkink et al. [GBVB21] als sie in ihrer recht neuen Arbeit versuchten die Hauptfaktoren für eine Niederlage im professionellen Bereich des belgischen Fußballs herauszufinden. Die Klassifizierung einer Niederlage gelang ihnen mit 89 Prozent in der Cross-Validation ziemlich erfolgreich, da geringe Torunterschiede und glückliche Siegestreffer eine Genauigkeit nahe den 100 Prozent nachezu unmöglich machen. Jedoch zeigt ihre Arbeit, dass eine hohe Genauigkeit (innerhalb der probabilistischen Rahmenbedingungen) möglich ist, und dass in Zukunft noch mehr im Detailgrad der Daten und ihrer Auffächerung geforscht werden sollte, um noch genauere Ergebnisse möglich zu machen.

Auch Jain et al. [JTS21] widmenten sich in ihrem Artikel aus 2021 der Vorhersage von Fußballergebnissen. In ihrer Arbeit verglichen sie die Leistung von zwei Machine-Learning-Ansätzen, nämlich den wiederkehrenden (= recurrent) neuronalen Netzen und den LSTMs. Sie nutzten den öffentlich zugänglichen Datensatz an Premier-League Matches aus den letzten Jahren als Basis, diese umfasst 300 Spiele und jeweils 60 Attribute. Es wurden dann für beide Modelle verschiedene Hyperparamerter, wie Batch-Größe, Chunk-Größe und auch verschieden große versteckte Schichten (hidden layer) getestet. Die Studie ergab, mit leichtem Overfitting, eine Trainingsgenauigkeit von 97-99 Prozent und eine Testgenauigkeit von 80-81. Eine davon ableitbare Konklusion ist, dass sehr viele Parameter den Ausgang eines Matches beeinflussen und somit die Vorhersage schwierig ist, jedoch eignen sich wiederkehrende neuronale Netze und auch LSTMs besser als herkömmliche Machine-Learning-Methoden für das Vorhersagen von Fußball Ergebnissen. Besonders die Leistung der LSTMs ist in diesem Bereich sehr vielversprechend und könnte in Zukunft auch in der Vorhersage von Spieler/innen-Leistungen Einsatz finden.

2.2.3 Tennis

Auch außerhalb von Wetten ist das Vorhersagen von Tennis-Matches ebenfalls von möglicher Bedeutung, da dadurch spannendere Matches zusammengestellt werden könnten und somit höhere Einschaltquoten, mehr Ticketverkäufe und somit mehr Umsatz möglich wäre [RP22]. Es wurden im Tennissport bereits einige Versuche unternommen, Ergebnisse vorherzusagen. Bereits 2005 wurde von Barnett und Clarke [BC05] analysiert, wie man mit öffentlich zugänglichen statistischen Daten der ATP Aufschlagstatistiken und in weiterer Folge auch Matchergebnisse vorhersagen kann. Sie wählten ein Markov Ketten Modell für ihre Arbeit aus und berechneten damit die dynamischen Wahrscheinlichkeiten, ob ein Spieler sein nächstes Aufschlag-Spiel gewinnen wird. Ihre Fallexperiment war ein Erfolg und sie propagieren, dass ihr Modell für alle Matches zweier ATP-Spieler anwendbar wäre.

Als weiteres Beispiel haben Bayram et al. [BGB21] ein Tennis-Match als Netzwerk

von Eigenschaften behandelt und sowohl Spieler/innen bezogene Eigenschaften, als auch die des Untergrundes und der Matchumgebung miteinbezogen. Insgesamt wurden ca. 80 Eigenschaften aus dem Datensatz entfernt. Dies wurde so gewählt, damit man eine genauere Aufschlüsselung der Leistungen auf den verschiedenen Turnieren und Untergründen (Belägen) bekommt, da die Beläge durch ihre großen Unterschiede in den Spieleigenschaften große Leistungsschwankungen in undifferenzierten Modellen sorgen. Folglich wurden mehrere herkömmliche und neumodische Machine-Learning-Ansätze verglichen. Es wurde klar, wenn auch nur um einen geringen Prozentsatz, dass neumodische Ansätze, wie Multi-Target Regression mit spezifischen Ziel-Eigenschaften (MTR-TSF) besser abschneiden, als traditionelle, wie beispielsweise lineare Regression (LR). Auch wurde in der Forschung innerhalb der traditionellen Machine-Learning-Methoden nach der beste Leistung gesucht. SVM, Zufalls-Wälder (random-forests = RF), Neuronale Netze und logistische Regression wurden gegenübergestellt. Sowohl der RF als auch die logistische Regression kamen auf eine Test-Genauigkeit von ca. 69 Prozent. Nur das neuronale Netz schnitt hier aufgrund von Overfitting auf die Trainingsdaten mit 65 Prozent etwas schlechter ab, was wahrscheinlich mit einem knappen Zeitrahmen und die damit verbundene Unmöglichkeit der Hyperparameter-Feineinstellung zusammenhängt [CSW17].

2019 wurde dann von Gao und Kowalczyk [GK19] eine groß angelegte Untersuchung durchgeführt, in der sie 24.594 Matches untersuchten, zusätzlich nutzten sie Informationen über Wettqouten für die Matches. Diese Informationen waren für 19.880 Matches aus dem Datensatz verfügbar. Dieser Datensatz wurde dann in drei verschiedene Machine-Learning-Ansätze eingespeist, die sich auch in den meisten bisher behandelten Studien wiederfinden, nämlich eine SVM mit radialem Basisfunktions-Kern, logistische Regression und RF-Klassifizierung. 18 Variablen wurden als relevant eingestuft und somit in den Modellen verwendet. Sie wurden sehr breit gefächert, um alle möglichen Einflüsse einfließen lassen zu können. Es wurden körperliche, psychische und umweltbezogene (platz-, match- und belagbezogene) Variablen gewählt, darunter wurden dann einzelne kombiniert, um verschiedene Leistungscharakteristiken der Spieler/innen zu repräsentieren. In den Resultaten war es interessant zu beobachten, dass Regression und auch die SVM am besten mit allen Variablen arbeiteten, die RF-Klassifizierung hingegen recht große Schwankungen aufweist, wenn man gewisse Variablen entfernt oder hinzufügt. Somit wurden einige Kombinationen getestet und die beste davon erreichte eine 80 prozentige Genauigkeit, was jedoch durch die recht geringe Gewissheit der Entscheidungen nach dem Endpunktestand nicht die beste Variante war. Die meisten Punkte (Kombination aus Genauigkeit und Gewissheit) erzielten die Vorhersagen basierend auf den Quoten, gefolgt von denen der SVM.

Regression ist generell stark in der Forschung vertreten und so wurde ein Modell aufbauend auf dem sogenannten Glicko-Algorithmus erstellt. Wie oftmals in Machine-Learning-Forschungen war das Vorverarbeiten und die Selektion von relevanten Eigenschaften ein recht großer Aufwand. Anschließend wurde dieses Modell mit anderen State-of-the-Art Algorithmen verglichen und behandelt. Es ist jedoch nicht so wahrscheinlich, dass

sich ohne menschliche Expertise diese Vorhersagen bewahrheiten. Deshalb sollten, bis deutlich größere Daten-Mengen zur Verfügung stehen und auch mehrere Variablen/-Parameter dokumentiert und auch für die Öffentlichkeit freigegeben werden, immer auch Expert/innen miteinbezogen werden [YCHH22]. In einer diesjährigen Arbeit von Randrianasolo und Pyeatt [RP22] wurde eine Analyse von WTA-Matches der Australian Open durchgeführt, wobei sie diverse Machine-Learning-Algorithmen und Spezifikationen verglichen haben. Die Genauigkeit ging nie signifikant über 70 Prozent hinaus, jedoch verzeichneten sie es als Erfolg, da nur wenig Trainingsdaten, nämlich 122 Matches verfügbar waren. Es wäre aber auch möglich, dass die zu geringe Anzahl an Parametern oder eine Kombination aus beidem dafür verantwortlich ist. Die höchste Genauigkeit wurde von einer lasso Regression erzielt, dicht gefolgt von zwei verschiedenen SVMs. Logischerweise erhöht sich dann die Wahrscheinlichkeit einer genauen Prognose während des Matches, wenn der Algorithmus Zugriff auf die aktuellen Spielstände und Daten bekommt. Interessant hierbei ist, wie gut sich Machine Learning auf gewisse Faktoren bereits einstellen kann, beispielsweise in Abbildung 2.2 (a), wo man sieht, wie sich nach ca. 240 gespielten Punkten das Momentum auf die Seite von Djokovic schlägt und die berechneten Siegwahrscheinlichkeiten von Federer abstürzen. In Abbildung 2.2 (b) sieht man dann hingegen die Siegeswahrscheinlichkeiten von Gulbis, der als Favorit gehandelt wurde. Das macht deutlich, wie stark das Modell auf Favoriten eingeht, vor allem am Anfang des Matches [LBM19].

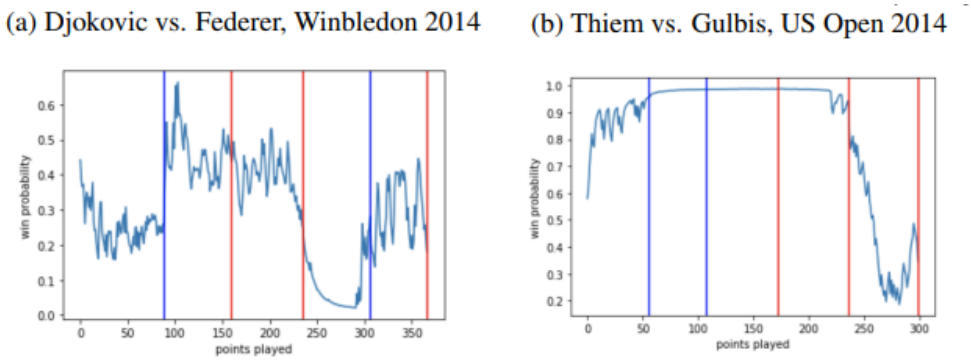

Abbildung 2.2: Exemplarische Darstellung wie sich Machine Learning auf Momentum-Umschwünge (a) und Favoriten-Rollen (b) einstellen kann. Bildquelle: [LBM19]

2.3 Vorgehensweise der Wettanbieter/innen

Der Markt der Sportwetten ist ein großer Teil des Sports, alleine in Deutschland wurden im Jahr 2012 € 3,59 Mrd. Spieleinsatzvolumen getätigt. Durchschnittlich setzen die Deutschen schätzungsweise 366 €, was jedoch stark nach demographischen Merkmalen variiert. Mit über 54 Prozent der platzierten Wetten ist Fußball mit großem Abstand der populärste Sport in der Wettbranche. Mit 17 Prozent folgt auf Platz zwei dann der Tennissport, gefolgt von Handball, Basketball und Eishockey mit jeweils ca. sechs Prozent [MadHAP14]. Wettanbieter/innen lassen sich immer von Buchmacher/innen sogenannte

Quoten ausrechnen, die die Wahrscheinlichkeit abbilden, ob das jeweilige Team oder der/die jeweilige Athlet/in gewinnen wird. Die Quoten starten von 1.0 und gehen in die Höhe je niedriger die Wahrscheinlichkeit eines Sieges ist. Zwar sind die Quoten dann eigentlich fixiert, jedoch gibt es, je nach Wettmarkt, auch Anpassungen, falls beispielsweise auf eine/n Mannschaft oder Spieler/in deutlich häufiger gewettet wird, so korrigieren die Buchmacher/innen die Quoten etwas nach unten, damit die Gegenoption für die Wettenden interessanter wird [GDBZ98]. Die Buchmacher/innen sind Expert/innen, auf deren Vorhersagen diese Quoten beruhen [SS09]. Wären die Quoten fair, so würde ihr Inverses die Wahrscheinlichkeit des erwarteten Gewinners darstellen, jedoch ist dies nicht der Fall, sondern die Quoten werden ein wenig niedriger angesetzt, sodass das Unternehmen damit sicher Gewinn macht, auch Spielraum genannt. Es gibt viele verschiedene Aspekte auf die gewettet werden kann [HŠŽ19]. Beispielsweise wird das Wetten auf eine gewisse Mindestanzahl/Maximalanzahl von Toren/Körben oder auch das Wetten auf ein Siegerteam mit einem gewissen Mindestvorsprung angeboten. Wird der Mindestvorsprung unterschritten, so gewinnt laut Wette das Verliererteam. [BDIG15]. Für diese Arbeit ist die relevanteste Variante das Wetten auf den/die Sieger/in des Wettkampfes, weshalb sie auch in den folgenden Unterkapiteln priorisiert wird. Aufgrund der Unterschiede in Genauigkeit, Rationalität der Quotenerstellung und Datenmenge werden [CGC01] sich die Unterkapitel jeweils auf den Wettmarkt einer Sportart fokussieren.

2.3.1 Fußball

Spann und Skiera [SS09] haben in ihrer Arbeit 2009 verschiedene Methoden zur Quotenerstellung und zur Vorhersage von Resultaten näher betrachtet. Die herkömmliche Variante sind die fixierten Quoten von Buchmacher/innen, die als Expert/innen ihre Erwartungen an den jeweiligen Wettkampf abgeben. Buchmacher/innen haben meist Zugriff auf relevante Informationen, die zusätzlich zu ihrer bereits vorhandenen Expertise und den allgemein zugänglichen Informationen der letzten Wettkämpfe, dabei helfen, sehr akkurate Vorhersagen treffen zu können.

Andererseits gibt es aber auch Ansätze, bei denen Quoten eher wie Aktien auf einer Börse gehandelt werden, sogenannte Vorhersage-Märkte (prediction markets = PM) [HŠŽ19]. Bei dieser Variante werden viele Personen über eine Plattform im Internet zusammengebracht und es wird mit den Ergebnissen, bzw. Gewinner/innen gehandelt und die Quoten ergeben sich, wie der Aktienkurs durch die Anzahl von Käufen und Verkäufen, und somit auch aus den gesamten Erwartungen der Teilnehmenden. Als dritte Variante, werden sogenannte Tipsters genannt, das sind einzelne Expert/innen, die in verschiedenen Zeitschriften und Magazinen ihre Erwartungen mit der Leser/innenschaft teilen. Diese Variante ist jedoch nicht von Relevanz, da ihre Genauigkeit sehr limitiert ist und somit nicht mit den anderen beiden mithalten kann. PMs und herkömmliche Wettquoten sind in ihrer Genauigkeit hingegen sehr ähnlich, jedoch ist es sogar möglich, dass die PMs bessere Ergebnisse liefern und somit, gäbe es nur moderate zusätzlichen Kosten, Profit gegenüber den Wettanbieter/innen machen könnten [SS09].

Diese PMs (auch bet exchange market genannt), wurden in den letzten Jahren rasch

beliebter. Franck et al. [FVN10] haben ebenfalls diese zwei Varianten der Wettmärkte gegenübergestellt mit einem großen Datenset von über 5000 Begegnungen aus fünf Ländern (Deutschland, Italien, Spanien, England und Frankreich). Ihre Ergebnisse legen ebenfalls nahe, dass die PMs eine höhere Genauigkeit erzielen als die Quoten der Buchmacher/innen. Plausible Mitgründe für diese Ergebnisse sind erstens, dass die versierteren Wetter/innen eher auf den weniger populären PMs wetten, als Laien, was natürlich die Genauigkeit der Gesamtleistung dieser Märkte steigert. Und zweitens werden die Quoten der Buchmacher/innen nochmals ein wenig angepasst, damit der Profit für das Unternehmen maximiert wird, was zu einer geringeren Genauigkeit führen könnte, da die PMs direkt die Meinung der Expert/innen übernehmen. Dennoch ist ersichtlich, dass die Quoten der Wettanbieter/innen eine gute Genauigkeit und Objektivität aufweisen, jedoch ein Machine-Learning-Algorithmus objektivere und rationalere Vorhersagen treffen kann [GS08]. Beispielsweise wurde bei den Matches ohne Zuschauer (Geisterspielen) in der Covid-19 Pandemie von den Buchmacher/innen eine deutlich niedrigere Anpassung des Heimvorteils in den Quoten vorgenommen als notwendig gewesen wäre. Dieser Umstand war dann folglich leicht für höhere Profite ausnutzbar für die Wettenden [FH22].

Ebenfalls wurde bereits versucht mit Machine Learning die Quoten der Buchmacher/innen zu schlagen, um damit Profit machen zu können. Jedoch ist ersichtlich, dass eine Angleichung und somit hohe Korrelation des Modells zu den Quoten der Buchmacher/innen einen geringeren Profit zur Folge hat, während Modelle mit geringerer Genauigkeit trotzdem mehr Profit erzielt haben, wenn eine Dekorrelation im Hinblick auf die gegebenen Quoten durchgeführt wurde. Mit dieser Herangehensweise und der Hilfe von gefalteten neuronalen Netzen wurden bereits 2019 Erfolge erzielt, sowohl in der Genauigkeit, als auch im hypothetischen Profit, der durch den Algorithmus erwirtschaftet werden könnte. Das belegt, dass Buchmacher/innen-Quoten nicht die definitive Lösung für Wettanbieter/innen darstellen und ein baldiger Umstieg auf Machine-Learning-Algorithmen durchaus denkbar wäre [HŠŽ19].

2.3.2 Basketball

Im College Basketball wurde eine Auswertung durchgeführt, die die Effizienz der Wettmärkte einzelner Teil-Ligen in Amerika (Conferences genannt) und etwaige Unterschiede in ihren Effizienzen erforschen sollte. Hierbei stellte sich heraus, dass recht große Unterschiede aufkommen in der Genauigkeit zwischen den verschiedenen Teil-Ligen. Dies lässt sich durch Informationsunterschiede durch verschieden ausgeprägte Berichterstattungen in Printmedien und Fernsehen erklären. Interessanterweise ist diese Schwankung bei der Effizienz der Wettanbieter/innen nicht merkbar [CGC01]. Da College Basketball, mit durchschnittlich ca. 5.600 platzierten Wetten pro Spiel (je nach medialer Aufmerksamkeit und Berichterstattung bis zu 34.000) [HPW13], viele Wettende anzieht, sind die Ergebnisse von Colquitt et al. [CGC01] ernst zu nehmen und aufgrund der Medienaufmerksamkeit auch recht gut mit dem professionellen Basketball in der NBA vergleichbar.

In der NBA ist das Wetten mit Wettanbieter/innen und Buchmacher/innen immer noch die deutlich beliebteste Variante. In dem Artikel von Gandar et al. [GDBZ98] wird

auf die Unterschiede zwischen Ausgangsquoten und Schlussquoten eingegangen. Diese unterscheiden sich durch die Einwirkung von Wettenden, die eine Anpassung der Quote erfordern, um die andere Option wieder profitabler zu gestalten. Die Frage war nun, ob diese Anpassung nur statistisches Rauschen ist. Jedoch zeigte die Untersuchung, dass die Anpassungen die Proportionen der Ausgangsquoten abbilden und ihre Auswirkung lässt die Quoten besser die zugrunde liegende statistische Ausgangssituation darstellen. Das führt auch dazu, dass die Vorhersagequalitäten der Quoten in diesem Prozess deutlich steigen. Diese Umstände legen es nahe, dass Personen mit mehr/besseren Informationen unter den Wettenden sind, da öffentliche Informationen alleine keine Vorhersagen, besser als die der Wettanbieter/innen, erlauben sollten.

Interessanterweise ist es ebenfalls möglich, dass verschiedene Wetten und ihre korrespondierenden Quoten die Wettleistungen von Wettenden verbessern können, indem sie möglicherweise zusätzliche Informationen beinhalten. Beispielsweise ist die Angabe einer Quote für die Wette, wie viele Körbe in dieser Begegnung geworfen werden, chancenerhöhend für Wettende, die auf den Sieg eines Teams setzen wollen. Dies geschieht deshalb, da diese Quoten zusätzliche relevante Informationen beinhalten, die ansonsten der Öffentlichkeit nicht bekannt wären und die somit zu einer besseren Vorhersageleistung führen. Das zeigt, wie limitierte Informationen die Chancen der Öffentlichkeit beim Wetten deutlich reduzieren und welchen Vorteil Wettanbieter/innen und Buchmacher/innen in diesem System haben [BDIG15]. Ebenfalls achten müssen die Buchmacher/innen auf eine Gefühls-Verzerrung (Sentiment Bias = SB), die eine stimmungsbezogene Auswahlbegründung bei den Wettenden ausdrückt. Die Forschung legt nahe, dass dieser SB von den Buchmacher/innen zu einem gewissen Grad in die Quoten miteinbezogen wird, jedoch nicht in einem Maße, dass damit ein großer Profit erzielt werden könnte. Dennoch würde das eine Ineffizienz belegen, die beispielsweise von einem Machine-Learning-Modell ausgenützt werden könnte [FHS18].

2.3.3 Tennis

Tennis ist, dadurch, dass es meist als Einzelsport ausgetragen wird (Doppel wird in dieser Arbeit vernachlässigt), sehr von den zwei Protagonist/innen des Matches und ihren psychischen und physischen Verfassungen abhängig. Danach ist es beispielsweise sehr leicht möglich, dass das psychologische Momentum zu einem ungünstigen Spielstand kippt und so ein völlig anderes Match entsteht und vielleicht der/die Sieger/in doch anders heißt [JM97]. Diese nicht leicht nachvollziehbaren psychologischen Schwankungen sind auch einer der Gründe dafür, dass Tennis sehr anfällig für Wettbetrug und Insider-Wetten ist. Sehr viele, mehr oder weniger bekannte Tennisspieler/innen sind bereits dieser Praktiken beschuldigt und zu beträchtlichen Teilen auch schuldig gesprochen worden. Die große Popularität des Tennissports und des Wettens auf Tennis-Matches in Kombination mit dem finanziellen System des Sports, das für Spieler/innen im Bereich 200 bis 800 der ATP/WTA Weltrangliste nur geringe Profite abwirft, sind zu großen Teilen ebenfalls an dieser Entwicklung beteiligt [RSG16]. Aber auch legales Insider-Wetten ist, je nach Definition, möglich. Insider kann nämlich so ausgelegt werden, dass Leute Informatio-

nen bekommen, bevor die allgemeine Wettgemeinde darauf Zugriff hat. Beispielsweise ist ein möglicher Grund für dieses verfrühte Wissen eine bessere Internet- oder auch TV-Verbindung. Ist das der Fall, werden Informationen der aktuellen Geschehnisse von anderen, in der Verbindung benachteiligten, Zuseher/innen erst später erhalten und verarbeitet. Somit entsteht ein Vorteil für die Vorhersage für jemanden mit schneller Verbindung, indem die Quoten noch nicht korrigiert worden sind [Bro12]. Oftmals gibt es auch kurzfristige Änderungen, wie die Aufgabe eines/einer Favoriten/Favoritin, die sich stark auf die Quoten der anderen Teilnehmer/innen auswirken, in dem Fall müssen Buchmacher/innen teilweise schnell reagieren und eine Lösung finden, die diese unvorhergesehene Änderung kompensiert. Am Beispiel von Wimbledon 2009 und dem Zurückziehen von Rafael Nadal sieht man, dass diese Änderungen weit über eine durchschnittliche Erhöhung der weiterhin im Bewerb stehenden Spieler/innen hinaus geht und individuell neue Quoten errechnet werden [LZH09].

Diesem Umstand sind sich Wettanbieter/innen und Buchmacher/innen natürlich bewusst und wie es scheint, gehen sie bereits erfolgreich dagegen vor [LF16]. Es zeigt sich aus den Quoten, dass der Spielraum, der zwischen den realen Wahrscheinlichkeiten und den tatsächlichen Wettquoten liegt, damit die Wettanbieter/innen ihren Profit garantieren können [HŠŽ19], über die Jahre sinkt. Das bedeutet, dass die Unternehmen Wege gefunden haben, ihren Profit anders zu sichern, sei es nun bessere Preisfestlegungsstrategien, niedrigere Kosten durch den größeren Anteil an Online-Wetten oder eben durch Sicherheitsmaßnahmen gegen Insider-Wetten. Dennoch wurde eine hohe Variabilität in den Spielräumen festgestellt, was darauf schließen lässt, dass die Buchmacher/innen verschiedene Risiken in verschiedenen Matches sehen und sich diese Spielräume speziell auf verschiedene match-spezifische und spielerbezogene Eigenschaften beziehen. Beispielsweise werden in Grand Slams durch ihre erhöhte mediale Aufmerksamkeit größere Mengen an Wettenden erreicht, was das Wettvolumen steigert, wodurch es möglich ist, bessere Quoten für die Wettenden anzubieten, ohne weniger Profit zu machen [LF16].

Auch müssen sich Wettanbieter/innen auf Wett-Regeln vorbereiten, die es einfach machen könnten, verschiedene Wetten mit öffentlich zugänglichen Informationen mit garantiertem Profit abzuschließen. Beispielsweise wenn die tatsächliche Chance, dass ein Spiel in einer bestimmten Phase des Matches von Spieler A gewonnen wird, höher ist, als die Quoten glauben, so muss man immer nur diese Regel anwenden, um in dieser Situation über die Zeit unumstößlich Gewinn zu erzielen. Diese Wett-Regeln sind aber im Tennis sehr selten und abhängig von dem Belag. Auf Rasen-Belägen gibt es die meisten Regeln, die durchschnittlich positive Resultate liefern, auf Hardcourt sind es schon deutlich weniger und auf Sand sind es nur noch drei. Das repräsentiert recht gut die hohe Sicherheit, mit der professionelle Tennisspieler/innen auf Rasen die Spiele gewinnen, in denen sie aufschlagen. Jedoch sind diese Regeln zu größten Teilen nicht statistisch signifikant. Dem folgt, dass der Wettmarkt im Tennissport effizent funktioniert [LV18].

Ein weiterer Grund für diese Effizienz ist eine Unwahrscheinlichkeits-Verzerrung (longshot bias = LSB), die im Tennissport vorhanden ist. Der positive LSB bedeutet, dass Favorit/innen und sichere Wetten (mit niedrigen Quoten) konstant vorteilhaftere Quoten

bieten, und somit langfristig besser für die Wettenden sind, als riskantere Wetten. Ein negativer LSB hingegen würde das Umgekehrte bedeuten, dies ist beispielsweise im Baseball-Wettmarkt beobachtbar [FM07]. Dieses Phänomen sollte ja eigentlich nicht existieren, da die Quoten die Wahrscheinlichkeiten eines Sieges des/der jeweiligen Spielers/Spielerin finanziell ausgleichen sollten, jedoch ist das in einigen Sportarten, so eben auch Tennis, häufig nicht der Fall. Indizien zeigen, dass der LSB wahrscheinlich ein Instrument ist, mit dem sich Wettanbieter/innen vor Insidern und anderen Risiken schützen. Beispielsweise wäre es auch denkbar, dass Personen bei Live-Wetten schneller reagieren als die Buchmacher/innen die Quoten an die neue Situation anpassen können und somit die Wettenden deutlich bessere Quoten erhalten würden als normal [Lah14].

Interessanterweise sieht man diesen Bias obwohl im Wettmarkt des Tennissports eigentlich einige Komponenten fehlen, die sein Aufkommen begünstigen würden, nämlich sind die Wettenden zu großen Teilen über den Sport gut informiert und die Transaktionskosten sind sehr gering. Es liegt somit nahe, dass es in verschiedenen Sportarten, verschiedene Gründe für den LSB gibt. Die Existenz des LSB ist also von mehreren Quellen belegt, jedoch wieso er auftritt ist noch nicht wirklich geklärt und bedarf noch weiterer Forschung. Dieser Umstand resultiert aus den spezifischen Merkmalen, die verschiedene Sportarten und somit auch ihre korrespondierenden Wettmärkte aufweisen [FM07]. Beispielsweise gibt es schon zwischen Damen- und Herrentennis in der höchsten Spielklasse (ATP und WTA) Unterschiede in der Erstellung der Quoten. Die Quoten sind beim Damentennis generell schlechter und bei unterschiedlichen Wettanbieter/innen findet man auch unterschiedliche Gründe für diese Geschlechterdifferenz. Grundsätzlich lässt sich ein Teil dieser Differenz auf geringere Informationszugriffsmöglichkeiten der Allgemeinheit durch geringere mediale Aufmerksamkeit erklären, jedoch gibt es auch Wettanbieter/innen, bei denen eine Bereinigung dieser Biases trotzdem relevante Differenzen in den Quoten ergibt. Diese Differenzen können nicht restlos erklärt werden und bedürfen weiterer Forschung [BCM21].

2.4 Visualisierung

Visualisierung ist die bildliche Darstellung von Sachverhalten. Ziel ist es somit, die Analyse, die Kommunikation und das Verständnis von diesen Sachverhalten zu verbessern, seien es nun Daten, Konzepte oder auch Modelle. Anhand dieses Ziels kann man auch Kriterien ableiten, welche wichtig für eine erfolgreiche Visualisierung sind. Darüber hinaus ist ebenfalls notwendig, dass sich der/die Betrachter/in den realen Kontext aus der jeweiligen Visualisierung ohne Probleme herleiten kann. Die Herausforderung ist es nun jedoch, die Daten so einfach wie möglich darzustellen, für Laien, aber auch so, dass für Experten eine genauere Analyse der Daten möglich ist, und somit Zusammenhänge erkannt werden können. Für diese unterschiedlichen Ziele, werden Visualisierungen nun in drei unterschiedliche Stufen eingeteilt. Erstens die explorative Analyse, hier werden die Daten unbehandelt erkundet, um daraus Information und Struktur herauszufiltern. In dieser Stufe gibt es noch keine Hypothesen. Diese gibt es erst in der nächsten Stufe, der konfirmativen Analyse. Hier geht es darum, Daten aus einem bestimmten

Blickwinkel zu betrachten, um eine Hypothese zu überprüfen. Die letzte Stufe ist dann die Präsentation, hier werden nun nur noch die Ergebnisse, die in den letzten beiden Schritten ausgearbeitet worden sind, ohne weiterem Erkenntnisgewinn für die Forscher der Öffentlichkeit präsentiert. Für alle diese drei Stufen gibt es je nach Themenbereich geeignetere visuelle Variablen, mit denen die Information dargestellt werden kann. Unter anderem kann man auf Variationen im Bereich Farbe, Größe, Form, Textur, Position, Textur, Tiefe und auch Bewegung zurückgreifen. Alle diese Variablen haben verschiedene Informationen, die sie besonders gut ausdrücken können, beispielsweise können Farbe sehr gut Unterschiede oder auch andere in den Farben kodierte Bedeutungen übermitteln (Rot = Gefahr, Grün = Sicher etc.) [SM13]. Im Falle dieser Arbeit sind ebenfalls Diagramme sehr wichtig, Diagramme haben auch spezifische Einsatzgebiete, in denen sie besonders gut die gewünschten Informationen übermitteln. Beispielsweise werden Balkendiagramme verwendet, um Unterschiede auszudrücken und Tortendiagramme haben ihren Einsatzbereich hauptsächlich in der Visualisierung von Verhältnissen [CMJ+20].

Durch den Zweck der Informationsübertragung und auch der Erforschung von Daten wird der Visualisierung ein großer Verdienst am Fortschreiten der Wissenschaft angerechnet und somit hat sich die Visualisierung immer weiterentwickelt. Es gibt bereits sehr viele Unterbereiche zur Visualisierung, die jeweils in gewissen Fachgebieten und Kontexten ihre Ziele am besten erreichen. Beispiele hierfür sind die Strömungsvisualisierung, die Volumenvisualisierung und auch die Strukturvisualisierung [VW05]. Letztere ist auch für Machine Learning sehr relevant, da die Black-Box Natur von einigen Machine Learning Ansätzen oft das Vertrauen in sie schwächt. Um diesem Umstand ein wenig entgegenzuwirken und das Verständnis für diese Algorithmen in der Gesellschaft zu steigern wird Strukturvisualisierung verwendet [XUD+19].

2.4.1 Visualisierung von Machine Learning

Da Machine-Learning-Methoden meist als Black-Box-Algorithmen gelten, ist es schwer nachzuvollziehen, welche Entscheidungen wie und wieso getroffen werden. Dieser Umstand ist für viele Menschen ein Grund, dem Machine Learning gegenüber skeptisch eingestellt zu sein und ihm nicht zu vertrauen. Deshalb ist es wichtig, dass man versucht, die verschiedenen Schritte, Abläufe, Parameter und Hyperparameter und ihre Auswirkungen auf die Klassifikation oder Vorhersage so gut es geht zu visualisieren [CMJ+20]. Hierfür ist erklärbare Künstliche Intelligenz (explainable AI = EAI) ein relevanter Begriff. Das ist ein Forschungsgebiet, das sich das Ziel gesetzt hat, Machine Learning transparent und erklärbar zu gestalten. Zwei Teilbereiche der EAI sind für die Forschung und somit die Aufschlüsselung von komplexen Machine-Learning-Ansätzen am relevantesten. Erstens das Konzept des transparenten Designs, was das Verstehen der Struktur und die Erklärung der einzelnen Komponenten als Ziel verfolgt und zweitens die Erklärbarkeit (= post-hoc explanation), die erschließen soll, wieso eine Machine Learning Variante auf ein gewisses Ergebnis kommt. Beide Bereiche sollen also die Nachvollziehbarkeit von Machine Learning erhöhen, wobei die Resultate ebenfalls als Visualisierungen dargestellt werden [XUD+19].

Computer-Vision, Gesundheit und Biologie, sind die Bereiche, mit den meisten Arbeiten in dem Fachgebiet der Visualisierung und die Multi-Class-Klassifizierung ist die beliebteste Ziel-Variable. Die Art der Visualisierung ist breit gefächert, jedoch meistens nur zweidimensional. Oft wird auf Farbe gesetzt, um Informationen in diese Grafiken einzubauen. Streudiagramme sind mit Abstand die Diagrammart, die am häufigsten verwendet wird, gefolgt von Balkendiagrammen, selbst erstellten Glyphen, Histogrammen und auch Liniendiagrammen [CMJ+20].

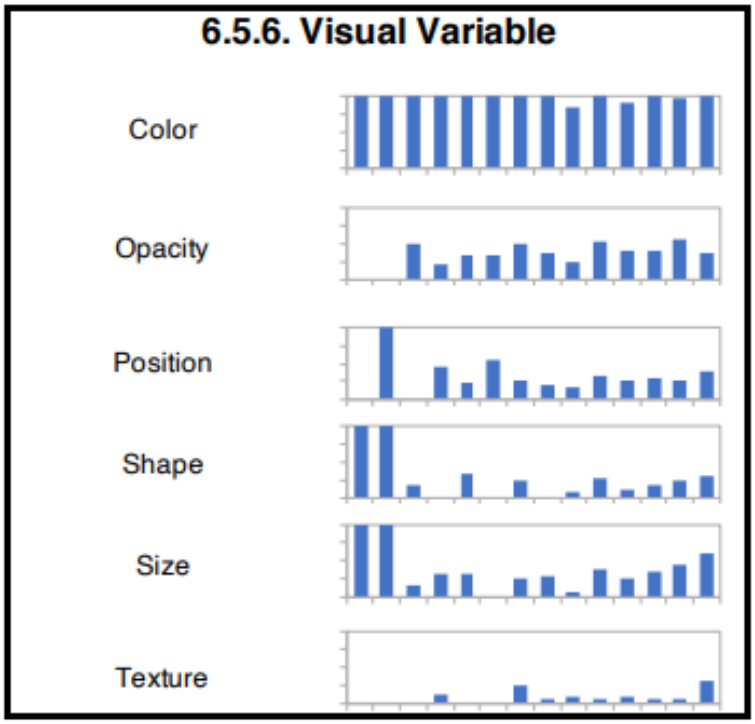

Abbildung 2.3: Beliebtheit der verschiedenen visuellen Darstellungsmethoden von 2007 bis 2020. Bildquelle: [CMJ+20]

Um jedoch die Architektur, beispielsweise die eines neuronalen Netzes darstellen zu können, wird meist ein ungerichteter/gerichteter Graph verwendet, da dies eine gute Verbildlichung der Arbeitsweise ergibt. Das Anpassen der Hyperparameter und die Auswirkung davon wird auch meist visualisiert, damit man auf einem Blick sieht, welche Vor- oder Nachteile sich aus den verschiedenen Auswirkungen auf das Resultat ergeben. Hierfür eignet sich ein Liniendiagramm sehr gut, da sich dadurch leicht, über einen zeitlichen Raum verteilte, Werte darstellen und auch bewerten lassen [MBR20].

Visualisierung von Ergebnissen einer Vorhersage wird meist verwendet, um klarer zu machen, was eigentlich das Ergebnis ist und es auch für so viele Menschen wie möglich zugänglich zu machen, aber auch um leichter Unterschiede, Trends und ähnliche Analysen daraus ableiten zu können. Für generelle Trend-Analysen eignen sich die Liniendiagramme sehr gut, da sie Unterschiede über die Zeit gut darstellen können, jedoch meist

für nur ein Team oder ein/e Spieler/in, da es sonst zu unübersichtlich wird [sin17]. Gibt es eine große Anzahl an einzelnen Einträge, die diskret klassifiziert werden sollen, so sind Streudiagramme sehr gefragt. Auch bei neuronalen Netzen kann die nicht-lineare Entscheidungsgrenze durch solche Diagramme in zwei oder mehreren Dimensionen abbildbar sein. In Abbildung 2.4 wird veranschaulicht, wie klassifizierte Daten, durch algorithmische Projektion (mit Hilfe des Deepview Algorithmus) von dem dreidimensionalen Raum, in den zweidimensionalen überführt wurden und dann auch noch die Entscheidungsgrenzen kalkuliert wurden. Natürlich ist diese Einfärbung der Entscheidungsgrenzen nur von Nutzen, wenn sich das Diagramm so nach Variablen aufteilen lässt, dass Ausprägungen, die in einem gewissen Areal liegen, nur zu einer der Klassen zugeordnet werden. Ist das nicht möglich, wie am Beispiel des praktischen Teils dieser Arbeit, so ist eine Visualisierung der Entscheidungsgrenzen nicht sinnvoll [SHH19].

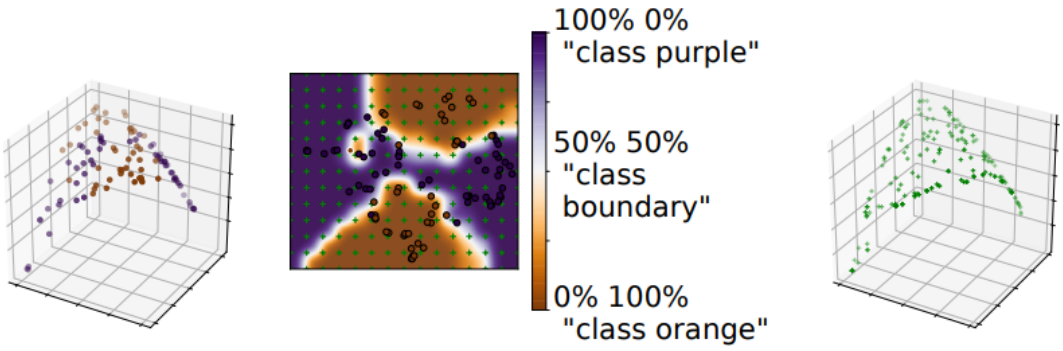

Abbildung 2.4: Visualisierung der Entscheidungsgrenzen eines 3D Klassifikators mithilfe des DeepView Algorithmus abgebildet auf 2D (Mitte) Wobei rechts die Daten in 3D und links die Ergebnisse der Klassifikation in 3D zu sehen sind. Bildquelle: [SHH19]

Visualisierungen von Machine Learning und den Ergebnissen ist für eine Analyse essentiell. Forscher/innen benötigen diese Verbildlichung, da Machine Learning meistens als Black-Box-Algorithmus auftritt und somit keine wirklichen Rückschlüsse zulässt. Deshalb ist es wichtig, die Ergebnisse so aufgeschlüsselt zu präsentieren, dass man daraus verschiedene Abhängigkeiten in den Parametern erkennen kann, die helfen die Entscheidung und die, ihr zugrunde liegende, Logik zu begreifen [KPN16]. Es werden häufig bereits interaktive Aspekte in den Visualisierungen eingebaut, da dies das Erkennen von Abhängigkeiten erleichtert. Seit 2007 waren nahezu alle Visualisierungen dieser Richtung mit der Funktion ausgestattet, einzelne Daten auswählen und gewählte Daten näher erkunden zu können. Darüber hinaus bietet eine beträchtliche Anzahl an Anwendungen die Funktion, nach beliebigen Daten filtern zu können, um den Blick auf gewisse Daten zu erleichtern. Auch die Funktion Daten zusammenfassen zu können, erwies sich als hilfreiche Interaktionsmöglichkeit [CMJ+20].

2.5 Frameworks und Libraries

Heutzutage ist es wichtig, dass Machine-Learning-Tools die Möglichkeit haben, über die Grafikkarte einen Teil der Rechenleistung zu erledigen, damit die Leistung und Geschwindigkeit gegeben ist, die notwendig/gewünscht ist. Hiervon werden in der Arbeit von Shi et al. [SWXC16] fünf der beliebtesten verglichen und analysiert, darunter Torch, Caffe, CNTK, MXNet und Tensorflow. Je nachdem, wie viele Threads laufen können, sind verschiedene Tools leistungsstärker. Caffe und Torch sind beispielsweise auf bis zu 4 Threads die effizientesten, bei höheren Threadanzahlen ist jedoch Tensorflow immer vorne. Jedoch gibt es in der Arbeit keinen Algorithmus, der die anderen behandelten Algorithmen, ungeachtet der Spezifikationen, immer leistungsmäßig übertrumpfen würde. Daraus folgt, dass es in erster Linie darauf ankommt, welches Computer-Setup man bieten kann, aber auch stark auf die Präferenz, welcher Algorithmus oder welches neuronale Netz nun genommen werden soll, weil man sich beispielsweise mit dem Tool auskennt [SWXC16]. Tensorflow ist auch die relevanteste Variante für diese Arbeit, es ist eine OpenSource Software Library von Google, die es erlaubt Machine Learning (auch Deep Learning) Modelle zu definieren, zu trainieren und auch anzuwenden [Gol16].

CHAPTER 3

Die Applikation

Im praktischen Teil dieser Arbeit wird eine Applikation zur Vorhersage von Tennis-Matches auf ATP Niveau mit Hilfe eines neuronalen Netzes vorgestellt. Das Ziel der Applikation ist es, vier neuronale Netze zu vergleichen, die Trainingsdaten aus unterschiedlichen Zeiträumen erhielten. Somit soll geprüft werden, welche Daten sich am besten für die Vorhersage eines Tennis-Matches eigenen, aber auch wie gut sich neuronale Netze generell für die Vorhersage des Gewinners in einem ATP Tennis-Match eignen. Darüber hinaus sollen die Ergebnisse für mehrere Spieler verglichen werden um so herauszufinden, ob Spiele, bei denen Spieler eines gewissen Niveaus beteiligt sind, sich besser oder schlechter vorhersagen lassen, aber auch, ob die Vorhersagen von der Zeit oder anderen Faktoren auch abhängig sind.

3.1 Datensatz

Der verwendete Datensatz umfasst 216.980 Einträge von Matches der ATP Tour [Nik20]. Der Fokus liegt hierbei auf Best-Of-Three Matches. Matches mit Best-Of-Five Wertung sind zwar auch enthalten, jedoch wurden diese auf die ersten drei gespielten Sätze reduziert. Der Datensatz umfasst 217 Spalten/Variablen, jeweils die gleiche Anzahl an Variablen für jeden der beiden Spieler, einige Informationen zum jeweiligen Match und eine zusätzliche für das Ergebnis dieses Matches, was in Folge der GT Wert für das Training der neuronalen Netze darstellen wird.

Die Variablen sind in vier Kategorien eingeteilt, die eine einfachere Übersicht und Zusammenfassung von Einflusskriterien erlauben. Es werden jeweils von beiden Spielern statistische Daten gelistet, die vor dem jeweiligen Match verfügbar sind. Diese Daten sind immer in der gleichen Kategorie, sie unterscheiden sich nur darin, dass verschiedene Zeiträume gewählt werden, um verschiedene Aspekte beleuchten zu können. Erstens werden Informationen der gesamten Profikarriere gespeichert, dazu wurden

Durchschnittswerte über diesen ganzen Zeitraum erfasst, die auch mit jedem nachfolgenden Match aktualisiert werden, mit diesen Daten kann eine grundsätzliche Einschätzung des jeweiligen Spielers erfolgen. Zweitens werden die Daten der vorherigen fünf Matches für beide Spieler gespeichert, diese Werte sind gut um einen Überblick zu erhalten, wie die Form des Spielers sich momentan verhält. Drittens werden diese Informationen auch zum letzten Match beider Spieler gespeichert, man bekommt somit eine aktuelle, aber sehr kurzfristige Momentaufnahme der beiden Spieler [Nik20].

Letztes Match	***Gesamte Karriere***	***Letzten 5 Matches***
Aces	Aces_A	Aces_L5
BreakPointsConvertedPCT	BreakPointsConvertedPCT_A	BreakPointsConvertedPCT_L5
BreakPointsConverted	BreakPointsConverted_A	BreakPointsConverted_L5
BreakPointsTotal	BreakPointsTotal_A	BreakPointsTotal_L5
DoubleFaults	DoubleFaults_A	DoubleFaults_L5
GRes	GRes_A	GRes_L5
IsHome	IsHome_A	IsHome_L5
K	K_A	K_L5
PS	PS_A	PS_L5
ReceivingPointsTotal	ReceivingPointsTotal_A	ReceivingPointsTotal_L5
ReceivingPointsWonPCT	ReceivingPointsWonPCT_A	ReceivingPointsWonPCT_L5
ReceivingPointsWon	ReceivingPointsWon_A	ReceivingPointsWon_L5
SETS_0-2	SETS_0-2_A	SETS_0-2_L5
SETS_1-2	SETS_1-2_A	SETS_1-2_L5
SETS_2-0	SETS_2-0_A	SETS_2-0_L5
SETS_2-1	SETS_2-1_A	SETS_2-1_L5
RID	-	-
Serve1stPCT	Serve1stPCT_A	Serve1stPCT_L5
Serve1stWonPCT	Serve1stWonPCT_A	Serve1stWonPCT_L5
Serve1stWon	Serve1stWon_A	Serve1stWon_L5
Serve1st	Serve1st_A	Serve1st_L5
Serve2ndWonPCT	Serve2ndWonPCT_A	Serve2ndWonPCT_L5
Serve2ndWon	Serve2ndWon_A	Serve2ndWon_L5
ServesTotal	ServesTotal_A	ServesTotal_L5
TPoints	TPoints_A	TPoints_L5
TTL	TTL_A	TTL_L5
TotalPointsWon	TotalPointsWon_A	TotalPointsWon_L5

Tabelle 3.1: Variablen des Datensatzes für jeweils das vorherige Match (), die ganze Karriere (_A) und über die letzten fünf Matches (_L5), diese Daten gibt es jeweils für beide Spieler (_1 und _2) [Nik20]

Wie in Tabelle 3.1 zu sehen, ist in dem Datensatz großer Wert auf die Aufschlagleistung der Kontrahenten gelegt worden. Es sind sowohl die Aspekte der gesamten Aufschläge und derer die im Spielfeld gelandet sind, beziehungsweise die erzielten Punkte, aber auch das Verhältnis zwischen gewonnenen Punkten/getroffenen Aufschlägen und gespielten Aufschlägen enthalten. Somit kann sichergestellt werden, dass eine geringe Aufschlagan-

zahl aufgrund eines kurzen und eindeutigen Matches nicht die Leistung des Aufschlägers verfälscht. Ähnlich verhält es sich bei den Break-Punkten (Break bedeutet, wenn ein Aufschlagspiel des Gegners oder der Gegnerin gewonnen wird), es werden sowohl absolute Anzahl, als auch Verhältnis angegeben. Auch hier ist beides für die Vorhersage relevant, da eine hundertprozentige Break-Punkt-Verwertung bei zwei Breakbällen im ganzen Match gegen eine fünfzigprozentige Verwertung bei acht Breakbällen verliert. Ebenfalls relevant sind die Ergebnisse, in Sätzen, der letzten Spiele, die Gesamtanzahl an gewonnenen Punkten (=TotalPointsWon) und auch Spiele (= TTL) und die Leistung als Rückschläger (= Returnspieler = Nicht-Aufschläger). Der Wert K beinhaltet dann noch zusätzlich die durchschnittlichen Quoten des Spielers und IsHome repräsentiert, ob die Spiele im Heimatland des Spielers stattgefunden haben [Nik20]. Das Feature RID ist das einzige, das nicht einheitlich in allen Kategorien vorhanden ist, es steht für die Runden-Id und bedeutet, welche Runde das gefragt Match war, somit ergibt dieses Feature nur für das vorherige Match (und das aktuelle Match) einen Sinn.

Nicht nur Daten aus den vorherigen Matches werden gespeichert, sondern auch Informationen zum aktuellen Match sind gelistet. Wie in Tabelle 3.2 ersichtlich, beschränken sich diese Daten auf Informationen, die bereits vor Spielbeginn öffentlich zugänglich sind und somit auch von einem Vorhersageprogramm, beziehungsweise Wettanbieter/innen und Buchmacher/innen verwendet werden können, um daraus Schlüsse zu ziehen.

Aktuelles Match
Age_CUR
DaysFromLast_CUR
HomeChanged_CUR
IsBirthDay_CUR
IsHome_CUR
IsLastRet_CUR
RID_CUR
Surface
SurfaceChanged_CUR
TourChanged_CUR
TourDist_CUR

Tabelle 3.2: Variablen des Datensatzes für jeweils das aktuelle Match (_CUR). Diese Daten gibt es jeweils für beide Spieler (_1 und _2) [Nik20]

Diese Informationen beinhalten Daten zu den Spielern, beispielsweise das Alter, ob einer davon gerade Geburtstag hat, wie lange das vorherige Match schon zurückliegt und auch ob dieses Match aufgegeben wurde. Zusätzlich gibt es noch Näheres zu dem Turnier, auf welchem Belag (Sand, Hartplatz oder Gras) das Turnier stattfindet, ebenfalls wie weit es von der Heimat der beiden Kontrahenten entfernt ist, ob es im Heimatland von einem der beiden ausgetragen wird oder auch ob sich für einen von ihnen der Belag oder die Turnierklasse im Vergleich zu dem vorherigen Match geändert hat [Nik20].

Zusätzlich zu den bereits genannten Eigenschaften umfasst der Datensatz noch weitere, beispielsweise gewonnene Punkte, Spiele und Sätze des aktuellen Spiels. Diese Informationen sind selbstredend nicht vor Beginn des Matches bekannt und somit auch nicht möglicher Teil eines Trainingsdatensatzes. Da das Ziel dieser Arbeit die Siegervorhersage ist, sind sie auch nicht als GT Werte relevant, in zukünftiger Recherche könnten jedoch die genannten Modelle auf die Vorhersage von exakten Endspielständen umfunktioniert werden. Dann gibt es noch einige Eckdaten zu dem Turnier, die nicht in die Modelle passen, jedoch teilweise für die Visualisierung dann relevant werden, wie beispielsweise Monat und Jahr des Matches. Andere dieser Daten wurden gar nicht verwendet, beispielsweise Turniername, Turnierland, Quoten für einen Sieg mit gewissem Vorsprung und so weiter, weil sie nicht für den Zweck dieser Arbeit relevant waren.

3.2 Methodik

3.2.1 Begriffserklärung

In diesem Kapitel wird erläutert, wie in dem praktischen Teil der Arbeit vorgegangen wurde. Da es hierbei um neuronale Netze und das Training dieser geht, gibt es einige Begriffe, deren Definition für dieses und auch die folgenden Kapitel essentiell ist. Ein ***Tensor*** ist ein Datentyp von Tensorflow [Gol16], der als Verallgemeinerung von Matrizen eingesetzt wird. Ein eindimensionaler Tensor ist ein Vektor, ein zweidimensionaler Tensor ist eine zweidimensionale Matrix, gleiches gilt für höherdimensionale Tensoren. Die ***Lernrate*** ist ein Hyperparameter der Machine Learning Trainings, der beschreibt, wie stark sich die Werte der Gewichte an die Informationen der vorherigen Iteration im Training eines Machine Learning Algorithmus anpassen sollen. Wird sie zu niedrig gewählt, so verbessert sich der Algorithmus nur langsam, wird sie zu hoch gewählt, gibt es große Schwankungen in der Genauigkeit des Algorithmus. Die ***Genauigkeit*** beschreibt, wie viele Daten des zur Verfügung gestellten Trainings- oder Testdatensets korrekt klassifiziert oder vorhergesagt werden, im Verhältnis zur Anzahl der gesamten Daten. Werden beispielsweise 100 von 200 Tennisspiele richtig vorhergesagt, so hat das Modell eine Genauigkeit von 50 Prozent. ***Loss*** hingegen beschreibt, wie stark sich die vorhergesagten/klassifizierten Werte von den GT Werten unterscheidet. Das bedeutet, Genauigkeit repräsentiert die Anzahl der Fehler und Loss repräsentiert die Schwere der Fehler, bzw. wie weit die vorhergesagten Werte von den wirklichen Werten entfernt sind [TOK18].

3.2.2 Die Modelle in Python

In diesem Kapitel werden die, für diese Arbeit, erstellen und trainierten Modelle beschrieben. Es wurde mit Tensorflow gearbeitet und es wurde ein "Keras Sequentiall Modell" [Ker22] ausgewählt. Dies bedeutet, dass ein neuronales Netz, mit definierbaren Schichtengrößen und einer definierbaren Schichtenanzahl als Basis fungiert. Dieses Modell der Keras API von Tensorflow ist ein einfacher und schneller Weg, sein neuronales Netz aufzubauen und zu trainieren. Jedoch ist es auch nicht ohne Nachteile, beispielsweise

ist es nicht für mehr als einen Eingabe- oder Ausgabetensor geeignet. Ebenfalls sind komplexere, nicht-lineare Architekturen nicht möglich. Für das Ziel dieser Arbeit ist das, und auch die weiteren Restriktionen dieses Modells für neuronale Netze, jedoch nicht relevant. Als Aktivierungsfunktion wurde eine sogenannte Sigmoid Funktion verwendet. Diese Funktion, die aufgrund ihrer Form auch S-Funktion genannt wird, beschreibt eine Variante, wie kalkuliert werden kann, ob ein Neuron nun aktiviert wird oder nicht. Dies geschieht dadurch, dass die Werte der eingehenden Gewichte in diese S-förmige Funktion eingesetzt werden, ist dieser Wert nun über einem Schwellwert (je nach Steigung), so feuert das Neuron, ansonsten feuert es nicht [Gol16].

Der Datensatz wurde in die bereits erwähnten fünf Kategorien (All / Last5 / Last / Current / nicht verwendet) aufgespalten, um daraus vier neuronale Netze zu modellieren, die jeweils mit einer der Kategorien (bis auf die Kategorie nicht verwendet) als Trainingsdaten arbeiten. Als Bibliothek wurde Tensorflow gewählt, aufgrund der guten Performance [SWXC16] und der guten Dokumentation. Außerdem wurde eine Webapplikation geplant und dafür ist die Funktion von Tensorflow sehr vorteilhaft, Modelle in Python trainieren zu können und sie dann in Tensorflow.js, der Javascript Variante von Tensorflow [Gol16], als umgewandeltes Json ohne großen Aufwand einbinden zu können. Somit wurde ein Modell mit den Daten für die gesamte Karriere der Spieler, eines mit denen zu den letzten fünf Matches, eines für das letzte Match und ein Modell für das aktuelle Match trainiert.

Die Modelle werden in den folgenden Kapiteln wie folgt bezeichnet: "All-Time-Modell", beruht auf den Daten der ganzen Karriere. Das "Last-5-Matches-Modell" beinhaltet die Daten für die letzten fünf Spiele. Das "Last-Match-Modell" klassifiziert aufgrund der Daten aus dem letzten Match und das "Current-Match-Modell" bedient sich an Daten zum aktuellen Match, die jedoch bereits vor Spielbeginn der Öffentlichkeit bekannt sind.

Auflistung 3.1: Codebeispiel: Einlesen der Daten, Aufbau und Training des Modells

```
data_last = pd.read_csv("/content/drive/MyDrive/Colab
Notebooks/atp.csv")

#Extrahiere die Ground Truth
data_last_labels =  data_last.pop('GRes_CUR_1')

#Entferne die Namen
data_last.pop("Name_1")
data_last.pop("Name_2")

#Beschreibe etweige Null-Werte der Datenquelle mit 0
data_last = data_last.fillna(0)

#Bereite die Trainingsdaten und -labels vor
data_last_train = np.array(data_last)
data_last_train_labels = np.array(data_last_labels)
```

```
data_last_train_labels = data_last_train_labels.astype(float)

#Definiere das Modell
optimizer1 = tf.keras.optimizers.Adam(clipvalue=0.0000000001)
model_last = tf.keras.models.Sequential()
model_last.add(tf.keras.layers.Dense(108, activation='relu',
                input_dim=54))
model_last.add(tf.keras.layers.Dense(1, activation='sigmoid'))
model_last.compile(optimizer=optimizer1, loss=
                'binary_crossentropy', metrics=['accuracy'])

#Trainiere das Modell
hist_last = model_last.fit(data_last_train,
                data_last_train_labels, epochs=50,
                batch_size=30, validation_split=0.2)

#Speichere das Modell als .h5 Datei ab
model_last.save("/content/drive/MyDrive/model_last_1.h5",
            save_format='h5')
```

In Auflistung 3.1 ist nun exemplarisch die Vorgehensweise beim Erstellen der Modelle in Python zu sehen. In einem ersten Schritt werden die Daten aus einer CSV Datei gelesen, die nur noch die, für das jeweilige Modell relevanten Features beinhaltet. Zusätzlich werden noch die Namen der beiden Spieler, für Testzwecke übergeben und natürlich ebenfalls die GT. Letztere ist die Information, welcher Spieler das Match der jeweiligen Zeile gewonnen hat. Zu finden ist die GT in der "GRes_CUR_1" Spalte. Diese wird entfernt und ein paar Zeilen darunter in die Form gebracht, in der Tensorflow sie benötigt, um mit ihr als GT arbeiten zu können. Dazwischen werden noch etwaige Null-Werte in der Tabelle mit Nullen (Integer 0) beschrieben, da Tensorflow ansonsten Probleme hat, diese Werte zu verarbeiten.

Im nächsten Schritt wird das Modell definiert. Die Ergebnisse des Modells hängen stark von der Konfiguration ab. Das Finden einer optimalen Konfiguration war in dieser Arbeit ein manueller Prozess. Interessanterweise war die Lernrate bis zu einem sehr niedrigen Wert von 0,0000000001 immer zu hoch, was zu instabilen Resultaten und auch Schwankungen in der Trainings-Genauigkeit führte. Die Größe des Hidden-Layers wurde nach längerer Optimierung auf das Doppelte der Input-Layer festgelegt. Das bedeutet für das Beispielmodell in Auflistung 3.1 108 Neuronen bei 54 Input-Variablen, während sowohl das Last-5-Matches-Modell als auch das All-Time-Modell 104 versteckte Neuronen bei 52 Eingangsvariablen aufweisen. Dieser Unterschied ist auf zwei zusätzliche Variablen im Last-Match-Modell zurückzuführen, nämlich die Runden-ID (RID) für beide Spieler, die in den anderen Modellen keinen Sinn ergeben würden. Das Current-Match-Modell liegt mit 20 Input Variablen und somit 40 versteckten Neuronen deutlich unter den anderen Modellen. Es wurde mit dem Keras Sequential-Modell gearbeitet

und als Aktivierungsfunktion wurde eine Sigmoid Funktion gewählt, da der Output entweder 0 oder 1 ist. Somit scheint Sigmoid und die binary_crossentropy die beste Hyperparameterwahl zu sein [Gol16]. Für das Training wurde dann ein angemessener Validierungssplit von 20 Prozent gewählt und durch die niedrige Lernrate werden 50 Epochen durchgeführt. Wie man in Abbildung 3.1 sehen kann, ist diese Epochenanzahl durch eine immer noch (wenn auch nur gering) steigende Genauigkeit und fallenden Loss gekennzeichnet. Durch die große Anzahl an Trainingsdaten wurde die Batchgröße auch so gewählt, dass eine ausreichend große Anzahl an Daten herangezogen werden kann, jedoch trotzdem das Training in einem angemessenen zeitlichen Rahmen durchgeführt werden kann. Zu guter Letzt werden die einzelnen Modelle dann noch als .h5 Dateien, also Dateien von verschachtelten Arrays mehrdimensionaler Art gefüllt mit wissenschaftlichen Daten [Dat22], abgespeichert, um innerhalb einer Python Umgebung in JSON-Dateien umgewandelt zu werden, die dann in Tensorflow.js geladen und verwendet werden können.

```
[ ] 5787/5787 [==============================] - 13s 2ms/step - loss: 0.6212 - accuracy: 0.6703 - val_loss: 0.7689 - val_accuracy: 0.6518
    Epoch 38/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6201 - accuracy: 0.6705 - val_loss: 0.7683 - val_accuracy: 0.6508
    Epoch 39/50
    5787/5787 [==============================] - 13s 2ms/step - loss: 0.6190 - accuracy: 0.6708 - val_loss: 0.7637 - val_accuracy: 0.6519
    Epoch 40/50
    5787/5787 [==============================] - 13s 2ms/step - loss: 0.6179 - accuracy: 0.6703 - val_loss: 0.7622 - val_accuracy: 0.6535
    Epoch 41/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6167 - accuracy: 0.6714 - val_loss: 0.7602 - val_accuracy: 0.6538
    Epoch 42/50
    5787/5787 [==============================] - 13s 2ms/step - loss: 0.6158 - accuracy: 0.6722 - val_loss: 0.7579 - val_accuracy: 0.6541
    Epoch 43/50
    5787/5787 [==============================] - 13s 2ms/step - loss: 0.6150 - accuracy: 0.6721 - val_loss: 0.7553 - val_accuracy: 0.6535
    Epoch 44/50
    5787/5787 [==============================] - 13s 2ms/step - loss: 0.6143 - accuracy: 0.6721 - val_loss: 0.7539 - val_accuracy: 0.6544
    Epoch 45/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6135 - accuracy: 0.6721 - val_loss: 0.7527 - val_accuracy: 0.6539
    Epoch 46/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6130 - accuracy: 0.6725 - val_loss: 0.7500 - val_accuracy: 0.6526
    Epoch 47/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6122 - accuracy: 0.6722 - val_loss: 0.7482 - val_accuracy: 0.6538
    Epoch 48/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6114 - accuracy: 0.6721 - val_loss: 0.7474 - val_accuracy: 0.6546
    Epoch 49/50
    5787/5787 [==============================] - 12s 2ms/step - loss: 0.6108 - accuracy: 0.6724 - val_loss: 0.7454 - val_accuracy: 0.6546
    Epoch 50/50
    5787/5787 [==============================] - 13s 2ms/step - loss: 0.6101 - accuracy: 0.6727 - val_loss: 0.7448 - val_accuracy: 0.6540
```

Abbildung 3.1: Letzte Epochen des Trainings. Es ist ersichtlich, wie der Loss (wenn auch nur noch gering) bis zur 50. Epoche sinkt und die Genauigkeit steigt.

.

3.2.3 Anwendung der Modelle in Tensorflow.js

Mit den fertigen JSON Dateien der trainierten Modelle musste nun entschieden werden, an welchen Spielern die Vorhersagefähigkeiten am besten getestet werden sollten. Aufgrund der bestehenden Möglichkeit, dass sich gewisse Trends in der Visualisierung abzeichnen, wurden Novak Djokovic und Rafael Nadal gewählt. Das Ziel war es, herauszufinden, was im Tennis am relvantesten ist, um regelmäßig zu gewinnen. Somit liegt es nur nahe, die zwei besten Spieler der letzten zwei Jahrzehnte auszuwählen. Ihr Spielstil ist sehr bekannt und durch die Visualisierung (die im nächsten Kapitel genauer behandelt wird) könnten womöglich Trends ersichtlich werden, die die Vorhersage leichter oder schwerer machen. Als dritter Spieler wurde jemand gewählt, der etwas außerhalb der aktuellen (Stand 19.06.2022) Top 100 der ATP Weltrangliste, nämlich auf Rang 169, steht, und zwar der slowakische Spieler Andrej Martin. Allein im Preisgeld sieht man den großen Unterschied

zwischen dem Erfolg von Djokovic (156.000.000 US Dollar) und Nadal (130.000.000 US Dollar) im Vergleich zu Martin (1.000.000 US Dollar), auch sieht man bei Nadal und Djokovic jeweils zirka 800 mehr Siege als Niederlagen auf ATP-Ebene. Martin dagegen hat 10 Niederlagen mehr als Siege. Matches in der Datenquelle sind von allen drei Spielern genügend vorhanden, auch wenn nur jene ausgewählt werden, in denen sie als Spieler 1 gelistet wurden. Für Novak Djokovic sind es 302 Matches, für Rafael Nadal 328 Matches und für Andrej Martin sind es mit 266 immer noch ausreichend, um eine Trendfeststellung zu erlauben. Somit ist ein Vergleich dieser Extremfälle sehr interessant und auch, ob sich diese Unterschiede im Erfolg in der Genauigkeit der Klassifizierung bemerkbar machen, oder ob sich vielleicht unter den Modellen einzelne herauskristallisieren, die für einen der drei Spieler besser oder schlechter funktioniert, als bei den anderen.

Diese Wahl sollte eine andere Perspektive zeigen und somit überprüfen, ob kategorische Unterschiede in der Vorhersageleistung auftreten, wenn man einen durchschnittlichen ATP Spieler mit zwei der erfolgreichsten Spielern aller Zeiten vergleicht. Es wurden zwei dieser Top-Spieler ausgewählt, da ihre Werte voraussichtlich eher Informationen liefern und eine Interpretation dieser im direkten Vergleich leichter fällt. Nach der Auswahl der Spieler wurde die Datenquelle so gefiltert, dass jeweils eine Datei mit allen Matches, in denen der jeweilige Spieler als Spieler 1 geführt wurde, angefertigt wurde. Dann wurden die Daten wieder in das richtige Format gebracht und die Modelle wurden initialisiert. Wie in Auflistung 3.2 ersichtlich, wurde dann eine Vorhersage Funktion geschrieben, die das jeweilige Modell auf die Daten des gewählten Spielers anwendet und ein Array aus richtig oder falsch klassifizierten Werten zurück gibt. Hierfür wird die model.predict() Funktion auf die in einen Tensor umgewandelten Datensätze des Spielers der Reihe nach angewandt, diese Funktion returniert einen Wert zwischen Null und Eins, der die Wahrscheinlichkeit repräsentiert, ob der jeweilige Spieler als Sieger aus dem Match hervorgegangen ist. Nun wird dieser Wert auf Null oder Eins abgebildet ($>0.5 \rightarrow 1$ / $<=0.5 \rightarrow 0$). Zusätzlich wird noch die Anzahl der korrekt klassifizierten Werte gespeichert, um die Auswertung einfacher zu gestalten und konkrete Prozentsätze der Genauigkeit errechnen zu können. Dieses Array mit den aus der Anzahl der korrekten Vorhersagen berechneten Daten über die Genauigkeit werden dann an die Visualisierungsfunktion übergeben, die die Ergebnisse optisch ansprechend und verständlich darstellen soll.

Auflistung 3.2: Funktion der Vorhersage in Tensorflow.js

```
function prediction(data, truth, model) {
    var result = [];
    for (let i = 0; i < data.length; i++) {
        const model_values = Object.values(data[i]);
        const val = model_values.map(Number)
        const a = tf.tensor([val]);
        const pred = model.predict(a);

        var readable_output = pred.dataSync();
        if (readable_output > 0.5) {
```

```
            if (truth[i] === '1') {
                count++;
                result[i] = 1;
            } else {
                result[i] = 0;
            }
        } else {
            if (truth[i] === '1') {
                result[i] = 0;
            } else {
                result[i] = 1;
                count++;
            }
        }
    }
    return result;
}
```

Die Architektur der Visualisierungsapplikation ist so aufgebaut, dass die Modelle zuerst geladen werden, dann an die Funktionen als Parameter weitergegeben werden, damit Redundanzen im Code vermieden werden können. Somit ist die Vorhersagefunktion aus Auflistung 3.2 die einzige Funktion, die dafür benötigt wird. Um das zu gewährleisten und auch damit die richtigen Parameter für das richtige Modell eingelesen werden, wurden Javascript Klassen für die Daten jedes Modells implementiert. Javascript Objekte besitzen die Eigenheit, dass nur existierende Variablen bei der Instanzierung einer Klasse befüllt werden. Deshalb wurden, nachdem die Daten mit der JQuery-CSV Bibliothek eingelesen wurden [JQu21], immer nur die Daten ins System übernommen, die auch für das jeweilige Modell relevant waren. In Tabelle 3.1 und Tabelle 3.2 ist ersichtlich, welche Variablen als Input für welches Modell notwendig sind, diese Variablen wären folglich auch jeweils notwendig, wenn künftig weitere Matches außerhalb dieser Datenbasis mit diesen Modellen erforscht werden sollen. Die CSV-Dateien an sich wurden aus einer, durch SQL Befehle gefilterten, Version der Ausgangsdatenbasis erstellt.

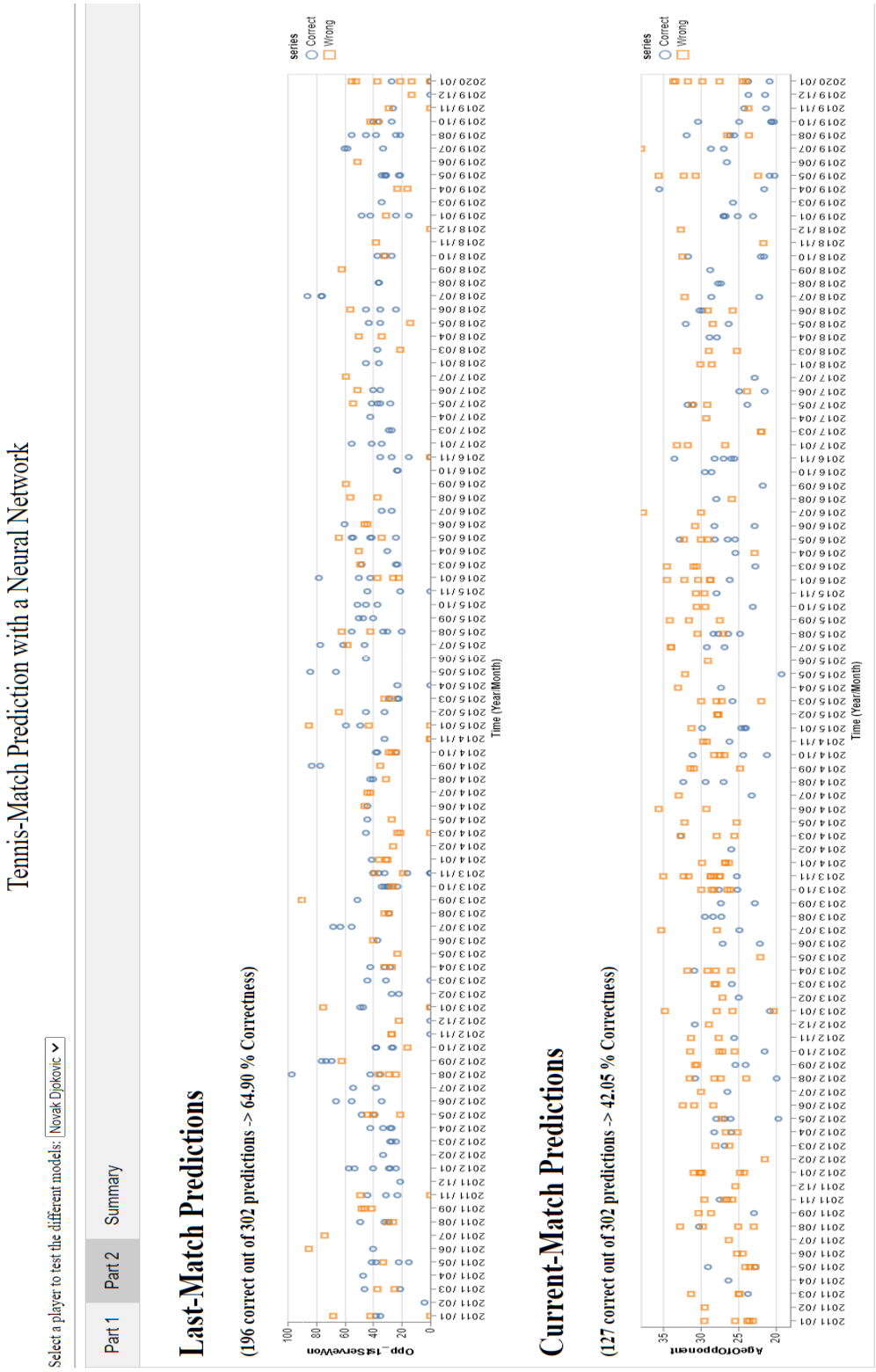

Abbildung 3.2: Gesamtansicht der Applikation am Beispiel des zweiten Tabs mit den Daten von Novak Djokovic. Die Y-Achse repräsentiert beim Last-Match-Modell die Punkte, die der jeweilige Gegner in seinem letzten Match gewinnen konnte und beim Current-Match-Modell das Alter des Gegners.

3.3 Visualisierung

Die Applikation wurde in der Entwicklungsumgebung Webstorm von Jetbrains [Web22] gestaltet und für die Visualisierung wurde die Tensorflow Erweiterung TFJS-VIS [TFJ22] verwendet. Diese Sub-Bibliothek macht es möglich, Tensoren und ähnliche Datentypen in Grafiken zu verpacken. TFJS-VIS verwendet Div-Elemente in HTML als Container, in die dann das gewünschte Diagramm hineingezeichnet werden kann. Zusätzlich kann durch TFJS-VIS nicht nur die einfache Ergebnis-Visualisierung, wie in dieser Arbeit, durchgeführt, sondern auch das Training und mehrere Aspekte des Modells an sich verbildlicht und kontrolliert werden. Das Ziel dieser Arbeit ist es, viele Informationen übersichtlich auf den Bildschirm zu zeichnen, ohne dass die Benutzer zwischen verschiedenen Ergebnissen wechseln müssen. Dazu wurde ein aus drei Tabs bestehendes Layout gewählt. Einmal wurden die zwei längerfristigen Modelle, also das All-Time-Modell und das Last-5-Matches-Modell, in einem Tab zusammengefasst, wobei in das zweite Tab dann die kurzfristigen Modelle, also das Last-Match-Modell und das Current-Match-Modell, eingebunden wurden. In dem dritten Tab wird eine Gegenüberstellung der einzelnen Genauigkeiten der einzelnen Modelle mit den Daten des ausgewählten Spielers angezeigt. Wie man in Abbildung 3.2 sehen kann, ist es möglich den Spieler über ein Dropdown-Feld links oben auszuwählen. Wird ein Spieler gewählt, so werden die Vorhersagen der Modelle neu berechnet, wodurch es je nach Rechenleistung eine gewisse Zeit dauern kann bis die Streudiagramme gezeichnet werden, da alle vier Modelle gleichzeitig mit der Auswahl ihre Vorhersagen starten und die Diagramme geladen werden.

Für die Darstellung der Modelle wurde ein Streudiagramm ausgewählt, mit zugleich farblicher und geometrischer Kodierung der zwei unterschiedlichen Datenreihen (richtig vorhergesagte Matches: blaue Kreise / falsch vorhergesagte Matches: orange Quadrate), wie in Abbildung 3.3 ersichtlich. Die X-Achse ist zeitlich von links nach rechts ansteigend in Monate gegliedert und erstreckt sich von Anfang 2011 bis Anfang 2020. Die Y-Achse ist je nach Variable und dem Vorkommen der Werte anders gegliedert, da sich die Werte aber ohnehin nicht unter den verschiedenen Modellen vergleichen lassen, ist dieser Umstand nicht von Relevanz für die Auswertung. Die Aufgliederung nach Monaten und die automatische Gliederung nach den Y-Werten gewährleistet auch eine gute Übersichtlichkeit in den Daten. Je nach Modell wurde für den Y-Wert ein ähnlicher, interessanter Wert verwendet, der etwaige Trends oder Ähnlichkeiten ersichtlich machen sollte. Bei dem All-Time-Modell, sowie auch bei dem Last-5-Matches-Modell und auch im Last-Match-Modell war es jeweils die Aufschlagstärke des Gegners in Form der gewonnenen Punkte bei erstem Aufschlag. Bei dem Current-Match-Modell hingegen war es das Alter des Gegners.

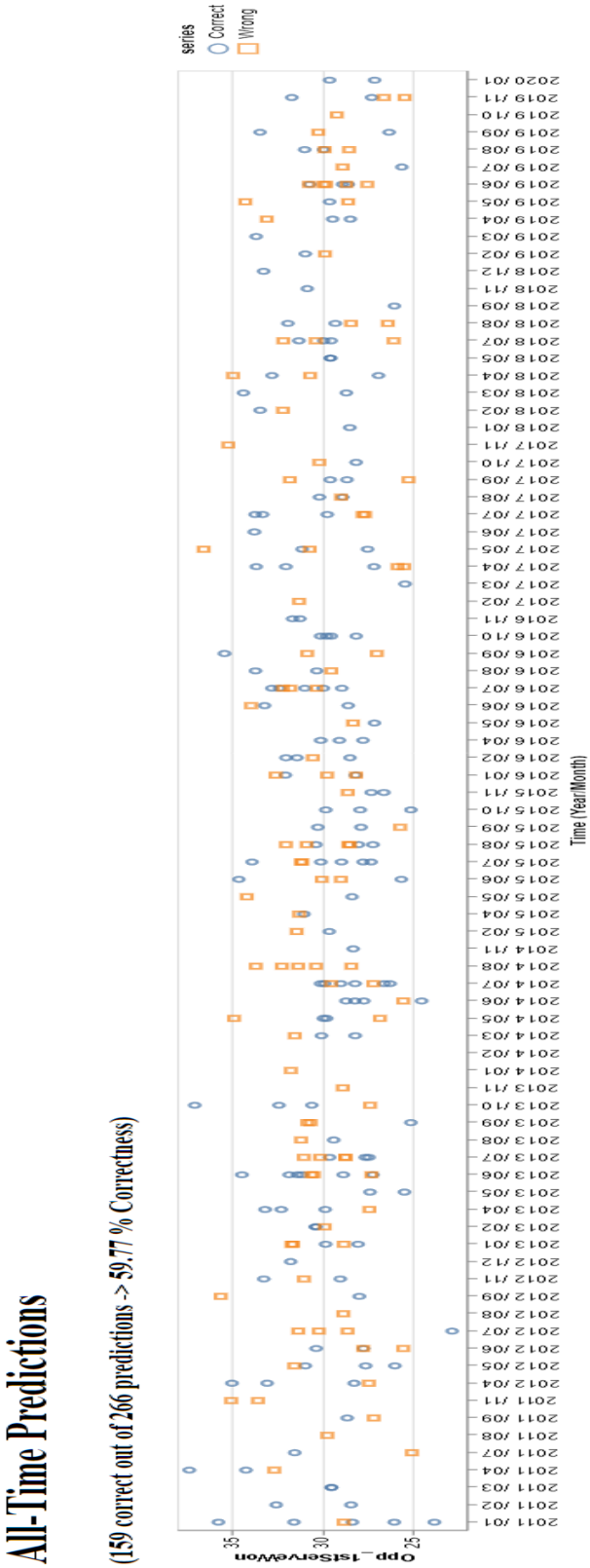

Abbildung 3.3: Exemplarisch ist hier das Streudiagramm des All-Time-Modells von Andrej Martin zu sehen.

.

Bei der Erstellung der Diagramme wurden die populärsten Visualisierungsmethoden in Betracht gezogen, sich deckend mit denen, die von Chatzimparmpas et al. [CMJ+20] recherchiert wurden. Somit wurden sowohl die Farbkodierung der verschiedenen Datenreihen eingesetzt, als auch die Form als zweiter visueller Kanal, damit auch sich überlappende Daten gut erkannt werden können und keine Information verloren gehen. Es wurden Streudiagramme zur Darstellung der Daten gewählt und es werden Interaktionsmöglichkeiten angeboten, die TFJS-VIS in den Diagrammen standardmäßig integriert, nämlich die Auswahl und nähere Inspektion einzelner Daten beim Darüberschweben mit der Maus. Wie illustriert in Abbildung 3.4 werden die Klassifizierung und die genauen Werte von Y-Achse und X-Achse nochmals angezeigt, was eine bessere Analyse und Gegenüberstellung der Daten ermöglicht.

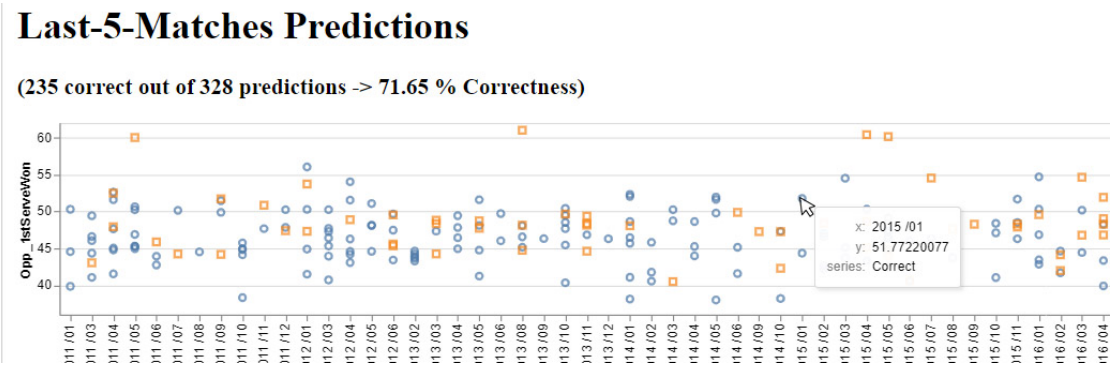

Abbildung 3.4: Auswahl eines Datums in dem Last-5-Matches-Modell von Rafael Nadal durch eine Mouse-Over Funktion.

.

So gut der Überblick über die Daten durch die Streudiagramme auch ist, so ist es nicht möglich, daraus konkrete Leistungswerte im Bereich der Vorhersagegenauigkeit der einzelnen Modelle herauszulesen, um einen stichhaltigen Vergleich durchführen zu können. Dazu wurde ein Diagramm als Zusammenfassung im dritten Tab eingefügt, das dieses Problem lösen soll. Für diese Zusammenfassung wurde ein Balkendiagramm mit gruppierten Balken gewählt, die wieder der gleichen Farbkodierung folgen, um Konsistenz zu bewahren. Dieses Diagramm wurde, wie die Streudiagramme zuvor ebenfalls mit TFJS-VIS implementiert. Folglich wird in jeder Gruppierung die Leistung eines Modells verglichen, auf der Y-Achse ist die absolute Anzahl richtig und auch falsch klassifizierter Spiele aufgetragen, wie ersichtlich in dem Beispiel von Novak Djokovic in Abbildung 3.5. Da Balkendiagramme sehr gut geeignet sind, um Unterschiede zu visualisieren und somit für vergleichende Nutzungszwecke ideal sind, können hier alle Informationen schnell und einfach abgelesen werden. Dieser Tab wurde dafür eingefügt, dass man Vergleichswerte zwischen den verschiedenen Modellen auf einen Blick abrufen kann, da dies mit den Streudiagrammen nicht möglich ist [CMJ+20]. Gemeinsam sollen alle Tabs alle relevanten Teilbereiche abdecken, um eine Einschätzung und Analyse der behandelten Modelle zur

Vorhersage des Siegers eines Herren Tennis-Matches zu ermöglichen.

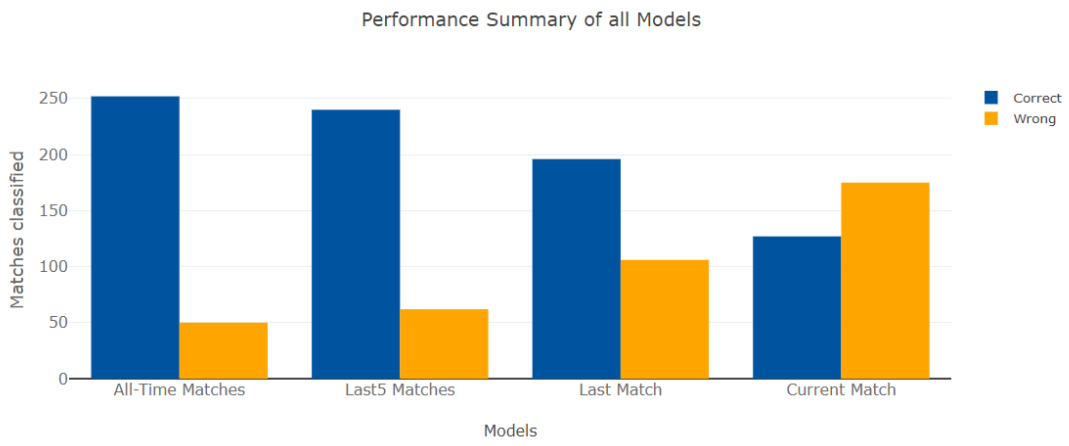

Abbildung 3.5: Zusammenfassung der Genauigkeit aller Modelle für Novak Djokovic.

.

3.4 Resultate

In diesem Abschnitt werden die Ergebnisse der Visualisierung der verschiedenen Modelle für den jeweiligen Spieler vorgestellt, und es wird ein Einblick in die Relation dieser Ergebnisse zueinander gegeben.

3.4.1 Resultate Novak Djokovic

Modell	Richtige Vorhersagen	Matchanzahl	Genauigkeit
All-Time-Modell	252	302	83.44 %
Last-5-Matches-Modell	240	302	79.47 %
Last-Match-Modell	196	302	64.90 %
Current-Match-Modell	127	302	42.05 %

Tabelle 3.3: Die genauen Daten der Modelle im Bezug auf die Vorhersage der Matches von Novak Djokovic

Novak Djokovic hat im Zeitraum zwischen Jänner 2011 und Jänner 2020 302 Matches gespielt, die für die vorgestellte Applikation herangezogen wurden. Es gab, wie in Abbildung 3.5 ersichtlich, recht große Unterschiede in den Leistungen der verschiedenen Modelle. Die beste Leistung erzielte das All-Time-Modell mit den Daten aus der gesamten Karriere des Spielers mit 252 richtig klassifizierten Matches, was einer Genauigkeit von 83,44 Prozent entspricht. Zweitbestes Modell war das Last-5-Matches-Modell, das mit 240 richtig erkannten Matches eine Genauigkeit von 79,47 Prozent erreichte. Das Last-Match-Modell reihte sich dahinter ein mit 196 richtigen Vorhersagen und 64,90 Prozent an Genauigkeit. Das mit großem Abstand schlechteste Modell war das Current-Match-Modell mit nur 127 korrekten Vorhersagen und somit einer Quote von 42,05 Prozent. Somit wurde in diesem Fall die Genauigkeit immer schlechter, je kurzfristiger die Daten wurden, wie zu sehen in Tabelle 3.3.

3.4.2 Resultate Rafael Nadal

Modell	***Richtige Vorhersagen***	***Matchanzahl***	***Genauigkeit***
All-Time-Modell	272	328	82.93 %
Last-5-Matches-Modell	235	328	71.65 %
Last-Match-Modell	202	328	61.59 %
Current-Match-Modell	168	328	51.22 %

Tabelle 3.4: Die genauen Daten der Modelle im Bezug auf die Vorhersage der Matches von Rafael Nadal

Rafael Nadal spielte in dem Zeitraum der gesammelten Daten 328 aufgezeichnete Matches, die für die Applikation ausgewählt wurden. Unter diesen Matches wurden von allen Modellen eine unterschiedliche Anzahl richtig vorhergesagt, wie in Tabelle 3.4 ersichtlich. Ähnlich wie bei Djokovic zuvor ist auch das All-Time-Modell das mit den meisten korrekten Vorhersagen, in diesem Fall 272, und mit einer Genauigkeit von 82,93 Prozent. Die Leistung des Last-5-Matches-Modells ist zwar ebenfalls noch auf Platz zwei, mit 235 richtigen Vorhersagen und mit einer Genauigkeit von 71,65 Prozent, jedoch im Vergleich prozentuell doch einiges schlechter als bei Djokovic. Das Last-Match-Modell ist nun wieder recht ähnlich mit 202 zutreffenden Klassifizierungen und somit einer Quote von 61,59 Prozent. Das Current-Match-Modell ist bei Nadal mit 168 richtigen Vorhersagen und somit 51,22 Prozent Genauigkeit um beinahe 10 Prozent genauer als bei Djokovic. Somit ist, wie bei Djokovic, ein (wenn auch nicht so starker) Abwärtstrend korrelierend mit der Langfristigkeit, also wie lang der Zeitraum ist, auf den sich die Daten beziehen, erkennbar. Je größer dieser ist, desto besser wird die Vorhersage. In Abbildung 3.6 ist dieser Trend sehr gut visuell beobachtbar, da die Werte für die korrekten Vorhersagen immer weiter sinken, je kurzfristiger die Daten werden. Bei den Daten der aktuellen Matches ist die Entscheidung und ihre Richtigkeit zirka 50 Prozent und somit gleichzusetzen mit einem Münzwurf.

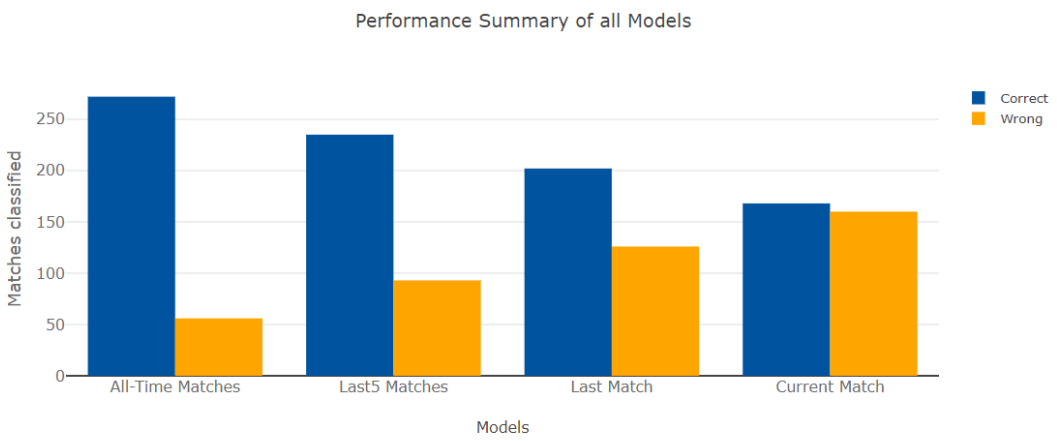

Abbildung 3.6: Zusammenfassung der Genauigkeit aller Modelle für Rafael Nadal.

.

3.4.3 Resultate Andrej Martin

Modell	Richtige Vorhersagen	Matchanzahl	Genauigkeit
All-Time-Modell	159	266	59.77 %
Last-5-Matches-Modell	154	266	57.89 %
Last-Match-Modell	135	266	50.75 %
Current-Match-Modell	140	266	52.63 %

Tabelle 3.5: Die genauen Daten der Modelle im Bezug auf die Vorhersage der Matches von Andrej Martin

Andrej Martin hat aufgrund seiner vielen Matches in Qualifikationen und tieferen Turnierstufen als der ATP, wie der Challenger-Tour und der ITF-Tour, weniger Matches in dem relevanten Zeitraum gespielt. Somit wurden 266 Matches herangezogen. Bei Martin gibt es die geringsten Unterschiede zwischen den Genauigkeiten der Modelle. Die Werte reichen von maximal 59,77 Prozent durch 159 richtige Klassifikationen beim All-Time-Modell, bis hin zu mindestens 135 korrekten Vorhersagen beim Last-Match-Modell, was 50,75 Prozent bedeutet. Dazwischen sind das Last-5-Matches-Modell, das mit 154 zutreffenden Klassifikationen und somit 57,89 Prozent knapp auf Platz zwei der Genauigkeiten liegt. Auf Platz drei ist schließlich das, bei Nadal und Djokovic immer am schlechtesten abschneidende, Current-Match-Modell mit 52,63 Prozent durch 140 richtige Prognosen. Somit sind die Leistungen der Modelle bei Martin mit Abstand am schlechtesten im Vergleich mit den anderen zwei Top-Spielern. Wie auch in Abbildung 3.7 gut erkennbar, sind die Balken meist sehr ähnlich hoch, was für eine hohe Zufalls-Quote spricht, auch die Modelle All-Time-Modell und Last-5-Matches-Modell, die bei den anderen beiden recht hohe Genauigkeiten erzielt haben, schaffen es nicht über eine 60 prozentige Genauigkeit hinaus.

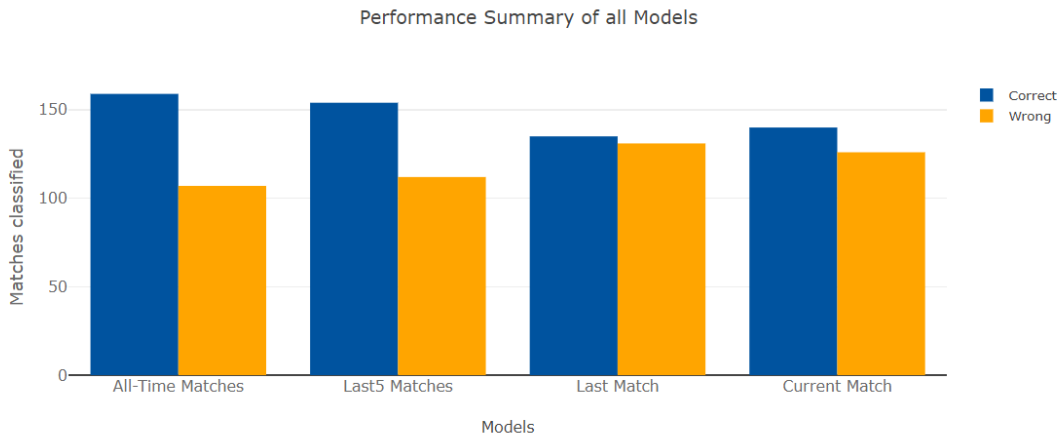

Abbildung 3.7: Zusammenfassung der Genauigkeit aller Modelle für Andrej Martin.

.

3.4.4 Resultate als Streudiagramme

Es wurden mehrere verschiedene Werte auf der Y-Achse getestet, doch einer der interessantesten Trends, die mutmaßlich vorhanden sind, ist die erhöhte Fehlklassifikation im Bereich von 45-50 gewonnenen Punkten des Gegners, wenn dieser den ersten Aufschlag trifft. Jedoch ist dieser Umstand nur in dem Last-5-Matches-Modell zu beobachten, wie in Abbildungen 3.9 und 3.8 ersichtlich. Ansonsten ist noch interessant, dass bei dem All-Time-Modell von Djokovic genau im letzten Monat der Aufzeichnungen fast alle Matches falsch klassifiziert wurden, ansonsten scheint es jedoch keine, über den Zeitraum der gespeicherten Matches (von Jänner 2011 bis Jänner 2020) zeitabhängige, Auffälligkeiten oder Trends zu geben. Nur bei dem Diagramm des Last-Match-Modells von Nadal scheinen mehr Werte falsch klassifiziert zu werden im zeitlichen Bereich zwischen 2013 und 2017. Außerdem sind mutmaßlich die meisten falsch vorhergesagten Matches von Nadal im Current-Match-Modell ab 23 Jahre alten Gegnern und nur bis Juli 2017, diese Einschätzungen beziehen sich auf die Modelle zu sehen in Abbildung 3.10.

Grundsätzlich scheint es so, als wären Auffälligkeiten mehr oder weniger auf die längerfristigen Modelle, also All-Time-Modell und Last-5-Matches-Modell beschränkt und die Werte der kürzerfristigen Modelle sind eher korrelationsfrei, beziehungsweise nur sehr schwach mit der gewählten Variable für die Y-Achse in Verbindung. Das kann auch in einem Zusammenhang mit der schlechteren Vorhersagegenauigkeit diese kurzfristigen Modelle stehen. Vor allem die Streudiagramme von Martin scheinen überhaupt keine Korrelationen zu beinhalten, wie in Abbildung 3.11 zu sehen ist, was ein weiteres Indiz für die schlechte Qualität der Vorhersagen bei den Matches von Andrej Martin ist.

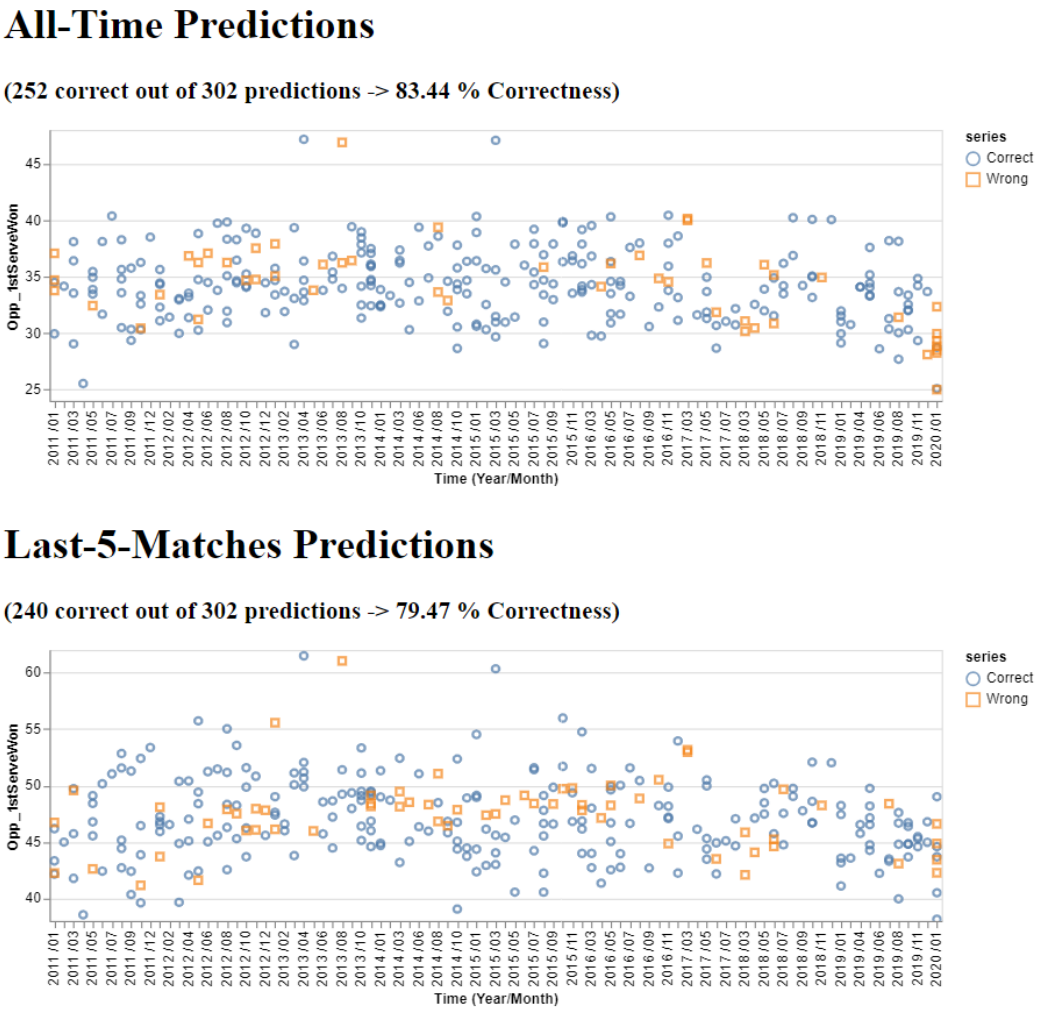

Abbildung 3.8: All-Time-Modell und Last-5-Matches-Modell Diagramme von Djokovic.

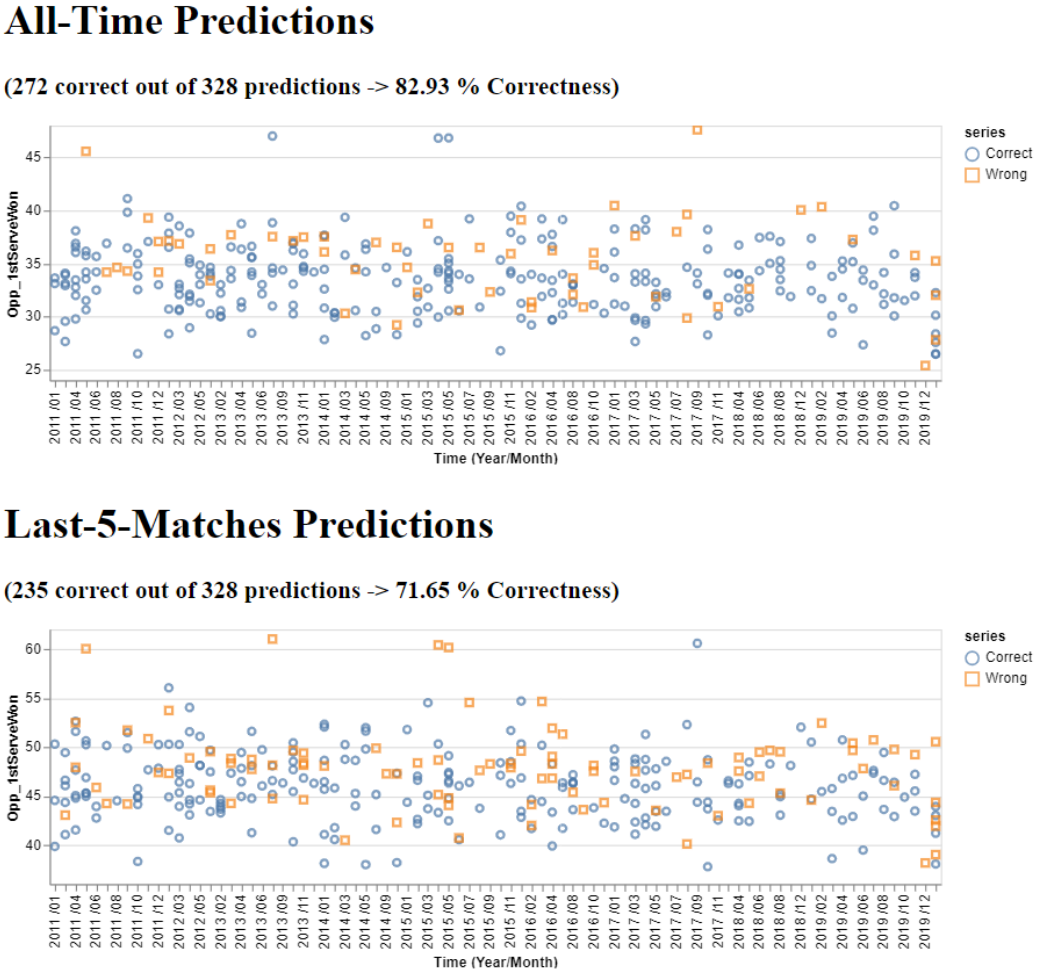

Abbildung 3.9: All-Time-Modell und Last-5-Matches-Modell Diagramme von Nadal.

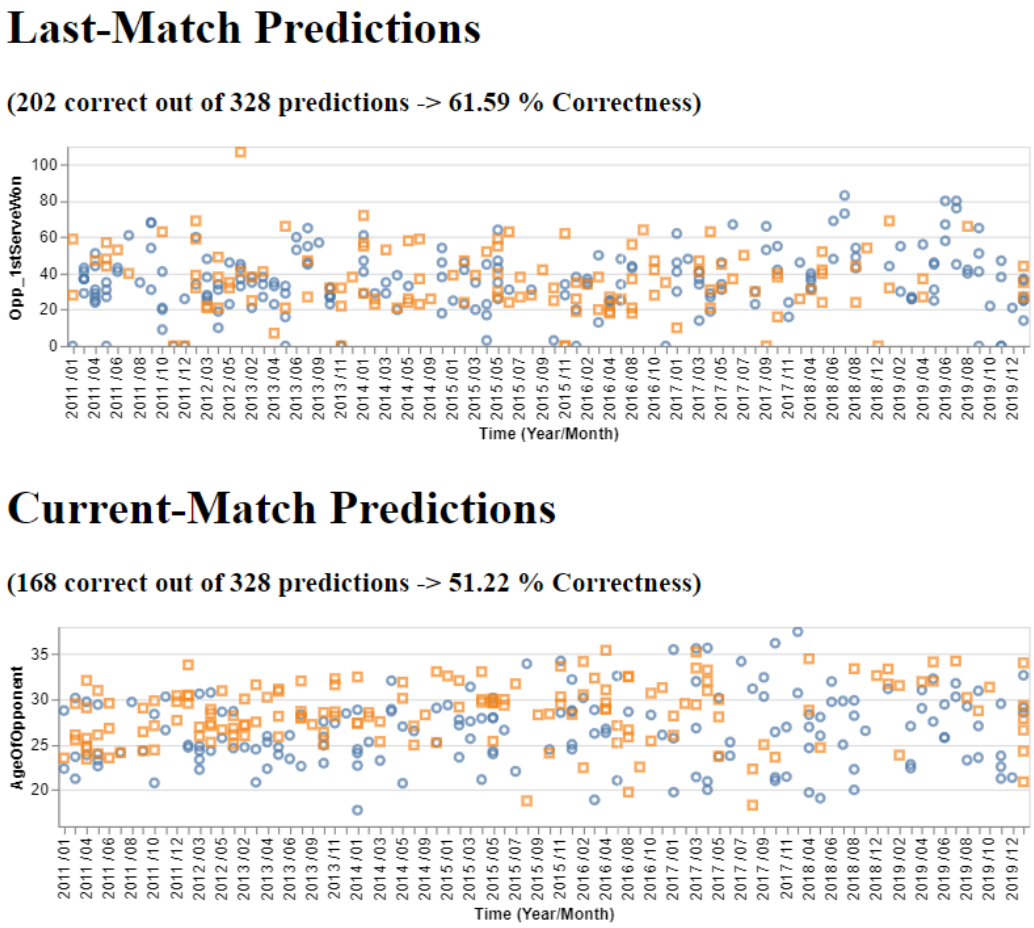

Abbildung 3.10: Streudiagramme der Modelle Last-Match-Modell und Current-Match-Modell von Nadal.

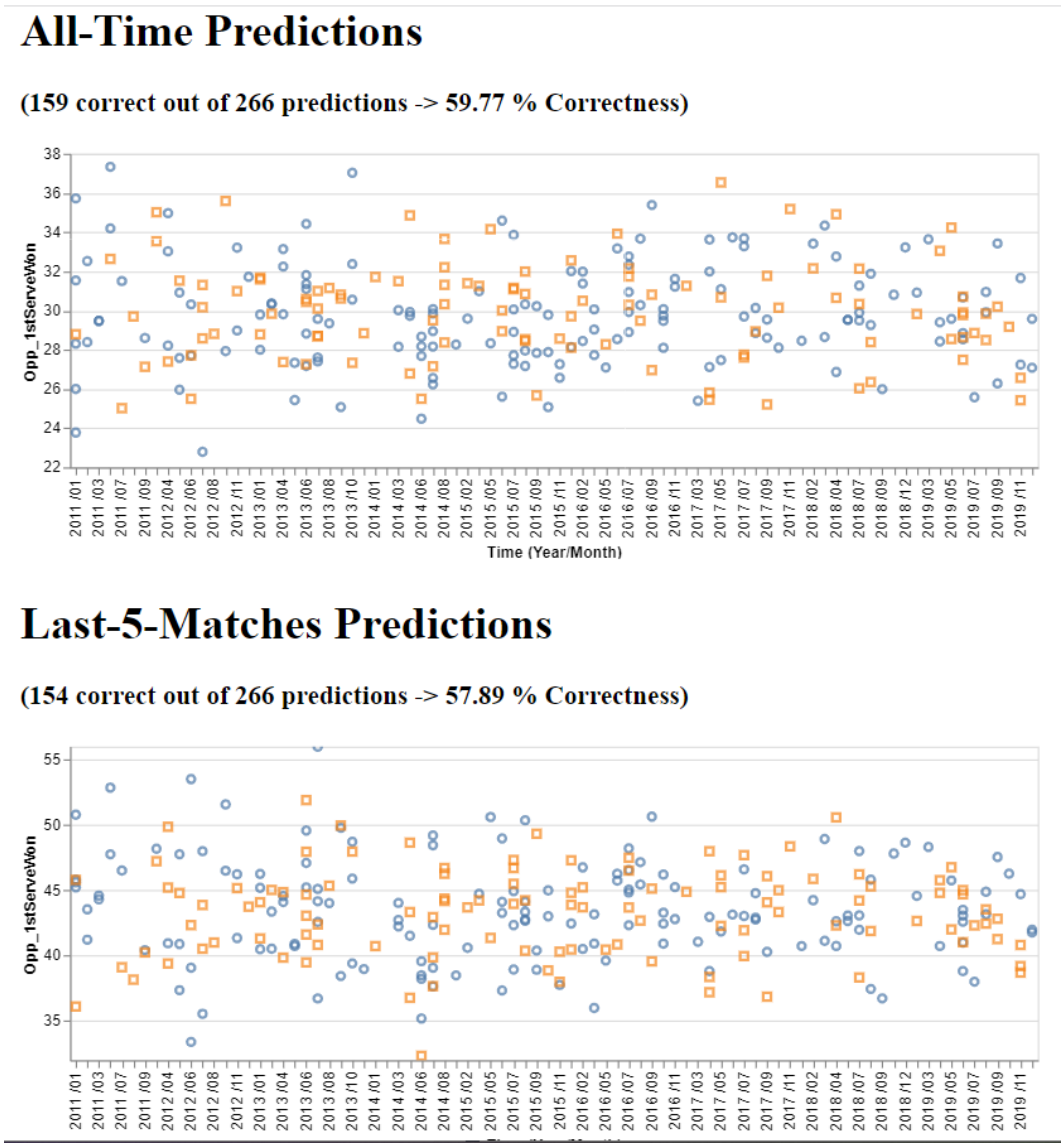

Abbildung 3.11: Streudiagramme der Modelle All-Time-Modell und Last-5-Matches-Modell von Martin.

CHAPTER 4

Diskussion

In diesem Abschnitt werden die Ergebnisse besprochen und mögliche Implikationen recherchiert. Allen voran werden die gefundenen Trends und Muster aus den Streudiagrammen und ihrer Zusammenfassung aufgegriffen und näher beleuchtet.

4.1 Einblicke und Schlussfolgerungen

4.1.1 Genauigkeiten der Modelle

Wie in den Abbildungen 3.5, 3.6, 3.7 ersichtlich, schneidet das All-Time-Modell immer am besten ab. Dies lässt die Vermutung zu, dass grundsätzliche Werte, die über die Karriere verteilt gesammelt wurden, mehr Einfluss auf die Leistung haben, als die Werte, die aus einer kleinen, dafür aber aktuelleren Zahl an Matches kommen. Somit wäre die Form eines Spielers, also die Leistungen der letzten Matches, statistisch nicht so hilfreich für eine gute Vorhersage, wie die generellen Leistungsmöglichkeiten. Das ist auch deckend mit den möglichen Leistungsschwankungen, die im Tennis aufgrund der mentalen Belastung und der psychischen Verfassung auftreten können [Grä03]. Diese Leistungsschwankungen würden nämlich bedeuten, dass die Leistung in den Matches über kurzfristige Intervalle möglicherweise stark fluktuiert. Sieht man sich jedoch die Statistiken der ATP von Djokovic und Nadal etwas genauer an, so erkennt man, dass bei ihnen diese Leistungsschwankungen eher selten auftreten und eigentlich meistens zumindest das Halbfinale oder Finale eines gespielten Turniers erreicht wird. Das lässt darauf schließen, dass die Schwankungen eher bei den Gegnern auftreten, was aufgrund der bereits genannten mentalen Herausforderungen des Tennissports logisch erscheint [Cow16]. Der Umstand, dass die Form des Spielers trotzdem nicht irrelevant ist, wird von dem, konstant am zweitbesten abschneidenden, Last-5-Matches-Modell gezeigt, was die Schlussfolgerung zulässt, dass die Daten des letzten Matches aus dem Last-Matches-Modell einfach zu kurzfristig sind, um wirklich gute Vorhersagen machen zu können. Das

könnte an den bereits erwähnten Leistungsschwankungen liegen. Jedoch liegt es auch nahe, dass die Vorhersagegenauigkeit hier am stärksten von den individuellen Eigenschaften des Spielers und der Situation abhängt. Diese These wird von den sehr großen Unterschieden zwischen den drei Spielern in der Genauigkeit des Last-Match-Modells gestärkt. Dadurch kommen bereits einige nicht bestimmbare Einflüsse hinzu, die die Vorhersage erschweren. Wohingegen das Last-5-Matches-Modell scheinbar bereits genügend langfristige Daten enthält, um auch diese Leistungsschwankungen besser abschätzen zu können. Dies wäre auch ein interessanter Bereich für weitergehende Forschungen, da man durch unterschiedliche Matchanzahlen in den Modellen erkennen könnnte, ab wann wirklich von einer Form des Spielers gesprochen werden kann und somit wie viele aktuellere Matches notwendig sind, um die beste Vorhersagegenauigkeit zu erzielen. Möglicherweise könnte dieses hypothetische Form-Modell dann sogar die Genauigkeit des All-Time-Modells überbieten. Das wäre nicht abwegig, da die Form eines/einer Spielers/Spielerin schwer zu erfassen ist, jedoch die Relevanz der Leistung der vorherigen Spiele ist unbestritten [Cow16], was auch diese Arbeit weiter bekräftigt. In der Literatur scheint die Form im Tennis noch nicht wirklich definiert zu sein, da die Recherche dazu erfolglos blieb.

Interessanterweise treffen diese Aussagen aber auf Martin nicht so stark zu wie auf die anderen beiden Spieler, dies könnte an der generell schlechteren Genauigkeit der Modelle bei den Matches von Martin liegen. Eine mögliche Erklärung für diese geringe Genauigkeit wäre die Kombination aus den Leistungsschwankungen beider Spieler, da diese Schwankungen bei Nadal und Djokovic, wie bereits erwähnt, recht gering sind, müssen mehr oder weniger nur die des Gegners berücksichtigt werden. Durch Martins eigenen, starken Schwankungen von Match zu Match, kommt ein weiterer Unsicherheitsfaktor zu der Vorhersage hinzu. Dieser Umstand würde auch erklären, wieso die Modelle alle nahe 50 Prozent Genauigkeit aufweisen, weil das neuronale Netz eigentlich nur raten kann, wer gewinnt. Ein weiterer Faktor, der in diese Unsicherheit hineinspielen könnte, ist, dass Martin auf verschiedenen Turnierebenen der ATP Tour spielt, somit gibt es eine deutlich größere Anzahl an verschiedenen Gegnern. Bei Nadal und Djokovic ist es eher selten, dass einer von ihnen gegen jemanden außerhalb der Top 100 der ATP Weltrangliste spielt, da diese Spieler sich nur selten für Turniere qualifizieren, in denen Nadal oder Djokovic antreten. Bei Martin, der auch auf der ITF Tour Matches bestreitet, sind die Gegner in einer deutlich größeren Bandbreite, das könnte es für das neuronale Netz schwieriger machen, da es weniger Referenzmatches in den Trainingsdaten gibt, in denen Martin gegen den jeweiligen Spieler gespielt hat. Wobei es bei Djokovic und Nadal meistens einige dieser Referenzmatches gibt, da zu größeren Teilen sich die Matches unter den besten Spielern wiederholen, beispielsweise gab es das Match Djokovic vs Nadal bereits 59 Mal. Durch die verschiedenen Häufigkeiten, wie oft gewisse Spieler gegeneinander antraten, würde die Genauigkeit eines Modells, das nur mit diesen Matches trainiert wurde, von Match zu Match sehr stark schwanken.

Es ist interessant, dass das Current-Match-Modell auch bei Nadal und Djokovic schlecht abschneidet, wobei es sowohl bei Nadal, als auch bei Martin den Anschein macht, als würde das neuronale Netz eher raten, da die Werte so nahe bei 50 Prozent sind. Bei

Djokovic sind es jedoch nur 42 Prozent, wie in Tabelle 3.3 sichtbar, was die Mutmaßung erlaubt, dass das Modell sehr wohl Informationen hat, diese jedoch falsch interpretiert und deshalb diese schlechte Genauigkeit resultiert. Dennoch fließen eigentlich mehrere, auf den ersten Blick relevant wirkende, Informationen in das Current-Match-Modell ein. Daten wie beispielsweise, ob das vorherige Match aufgegeben wurde, wie lang das letzte Match her ist, oder auch ob sich der Belag seit dem geändert hat, können leicht einen Unterschied ausmachen, wie an dem aktuellen Beispiel der Rückkehr von Dominic Thiem 2022 nach längerer Pause erkennbar. Das erste Match danach verlor er gegen jemanden, der über 200 Plätze in der Weltrangliste hinter ihm war. Selbst die Wettanbieter/innen sahen ihn, mit einer Quote im Bereich von 1.14 - 1.25, klar zu favorisierten [Cac22]. Somit ist es nicht unrealistisch, dieser Pause einen hohen Einfluss auf die Niederlage zuzuschreiben.

Die geringe Genauigkeit könnte unter anderem aus einer zu geringen Anzahl an Matches, in denen diese Ausnahmefälle eintreten, resultieren. Zusätzlich ist es auch wahrscheinlich, dass die verwendeten 20 Eigenschaften, die von dem Current-Match-Modell verwendet werden, alleine zu geringen Informationsgehalt beinhalten und deshalb nicht für eine Klassifizierung ausreichen. Im Vergleich zu den anderen Modellen mit jeweils 52, 52 und 54 Variablen ist dieses neuronale Netz natürlich sehr klein und einfach gehalten, jedoch gibt es auch nicht viele andere Variablen zum aktuellen Match, die öffentlich zugänglich sind und kein Expertenwissen, wie das von Buchmacher/innen, erfordern [SS09]. Diese Abtrennung war dem Aufbau der Arbeit und dem Vergleich der Modelle anhand ihrer Daten geschuldet. Für eine maximale Vorhersagegenauigkeit müssten diese Werte des aktuellen Spiels mutmaßlich mit den Parametern der anderen Modelle jeweils kombiniert werden. Wie in dieser Arbeit gezeigt, sind diese Werte alleine nicht aussagekräftig genug, um auf ihrer Basis erfolgreich den Ausgang eines Matches vorherzusagen.

4.1.2 Streudiagramme

In den Streudiagrammen sind nur geringe Korrelationen ersichtlich, was möglicherweise dem gewählten Y-Wert geschuldet sein könnte. Jedoch wurde dieser gewählt, da Aufschlag und Return im Tennis die relevantesten Schläge sind. Diese Schläge entscheiden, wer ein Match gewinnt und wer verliert, sogar auf Sandplatz, dem langsamsten Belag [GLTS09]. Somit war der Hintergrund dieser Wahl, dass man mit den gewonnenen Punkten des Gegners beim Aufschlag des Gegners erkennen kann, wie souverän die Aufschlagleistung war. Folglich wird geprüft, wie stark die Aufschlagleistung und Siegwahrscheinlichkeit korrelieren. Interessanterweise zeigt sich bei den Diagrammen der Last-5-Matches-Modellen ein Trend in der Fehlererkennung sowohl bei Djokovic als auch bei Nadal (wenn auch ein wenig geringer), was bestärkend für diese Hypothese wirkt.

Wie in Abbildungen 3.9 und 3.8 ersichtlich, liegt eine deutlich überproportionale Menge an falsch vorhergesagten Matches im Bereich 45-50 gewonnener Punkte des Gegners bei seinem ersten Aufschlag. Das würde die Schlussfolgerung erlauben, dass hier ein Bereich auftritt, in dem nicht ganz klar ist, ob durchschnittlich genug Punkte gemacht werden, oder nicht, beziehungsweise ob die Aufschlagleistung über die letzten fünf Matches

durchschnittlich gut genug war, um das jeweilige Match zu gewinnen, oder nicht. Somit wäre denkbar, dass das neuronale Netz aufgrund der mittelfristigen Aufschlagleistung aus den letzten 5 Matches, zu welcher es mehrere Parameter in dem Datensatz gibt, grundsätzlich gut unterscheiden kann, ob jemand gewinnt oder verliert. Nur in dieser Zone von 40-50 gewonnenen Punkten herrscht eine gewisse Unsicherheit und somit werden die Matches zu einem deutlich überproportionalen Teil, im Vergleich zu den restlichen Matches, falsch vorhergesagt.

Im All-Time-Modell ist dieser Trend nicht wirklich erkennbar, was mutmaßen lässt, dass möglicherweise die Aufschlagleistung auf die gesamte Karriere gerechnet genauere Vorhersagen erlaubt und somit die falsch klassifizierten Werte aufgrund einer anderen Variablen auftreten. Eine weitere Auffälligkeit liegt sowohl in dem All-Time Modell als auch in dem Last-5-Matches Modell von Djokovic, die aus nicht nachvollziehbaren Gründen beinahe alle Matches im Jänner 2020 falsch klassifizierten, was durch die hohe Genauigkeit beider Modelle und der ansonsten großen Streuung der falsch vorhergesagten Matches auffällig erscheint. Nach einer Recherche zu diesen Matches ist es noch unerklärlicher, da Djokovic diese Matches alle gewann und er gegen niemanden spielte, den er zuvor nie geschlagen hätte. Auch die letzten fünf Spiele vor den jeweiligen Matches waren nicht von großen Überraschungen geprägt.

Somit bleibt diese Entwicklung mit den Modellen und der Herangehensweise dieser Arbeit nicht erklärbar und bedarf weiterer Forschung. Vergleicht man nun alle All-Time-Modelle im Bereich von 30-40 Punkten bei eigenem Aufschlag des Gegners der drei Spieler untereinander, ersichtlich in Abbildung 4.1, so stellt man fest, dass sich die Gegner von Nadal und Djokovic in ihrer Aufschlagleistung über ihre Karriere gerechnet sehr stark von den Gegnern von Martin unterscheiden. Sichtlich mehr Matches von Djokovic und Nadal, in der Zeitspanne des Datensatz, wurden gegen Gegner ausgefochten, die durchschnittlich in ihrer Karriere mehr Punkte beim eigenen ersten Aufschlag machten als die Gegner von Martin. Bei Nadal und Djokovic ist die überwiegende Mehrheit der Gegner im Bereich von 30-40 Punkten bei eigenem ersten Aufschlag zu finden. Wogegen die Gegner von Martin in der Mehrheit bei 26-32 Punkten liegen. Das unterstützt wiederum die Resultate, dass Aufschlag und Return die wichtigsten Schläge sind und dass eine hohe Punktequote beim ersten Aufschlag essentiell ist, um zu den Besten im Tennissport zu gehören [GLTS09].

Abschließend sei noch ein nicht ganz so auffälliger Trend im Current-Match-Modell von Rafael Nadal erwähnt, ersichtlich in Abbildung 3.10. Aufgrund der sehr geringen Genauigkeit von 51 Prozent schien es sehr unwahrscheinlich, Trends zu erkennen, jedoch wirkt es, als würde das Modell in der ersten Hälfte, also bis 2015, deutlich mehr Matches falsch klassifizieren als in der zweiten Hälfte. Zusätzlich werden, je weiter nach rechts, in Richtung 2020, gegangen wird, immer mehr Matches gegen immer ältere Gegner besser erkannt. Die Ursache dieser Auffälligkeit ist schwer bestimmbar, jedoch ist eine plausible Teilerklärung, die Relevanz der Merkmale wie "Pause seit dem letzten Match" und "wurde das letzte Match aufgegeben?". Diese These ist deshalb wahrscheinlich, da er seit 2012 sehr häufig mit verschiedensten Verletzungen zu kämpfen hatte und somit möglicherweise

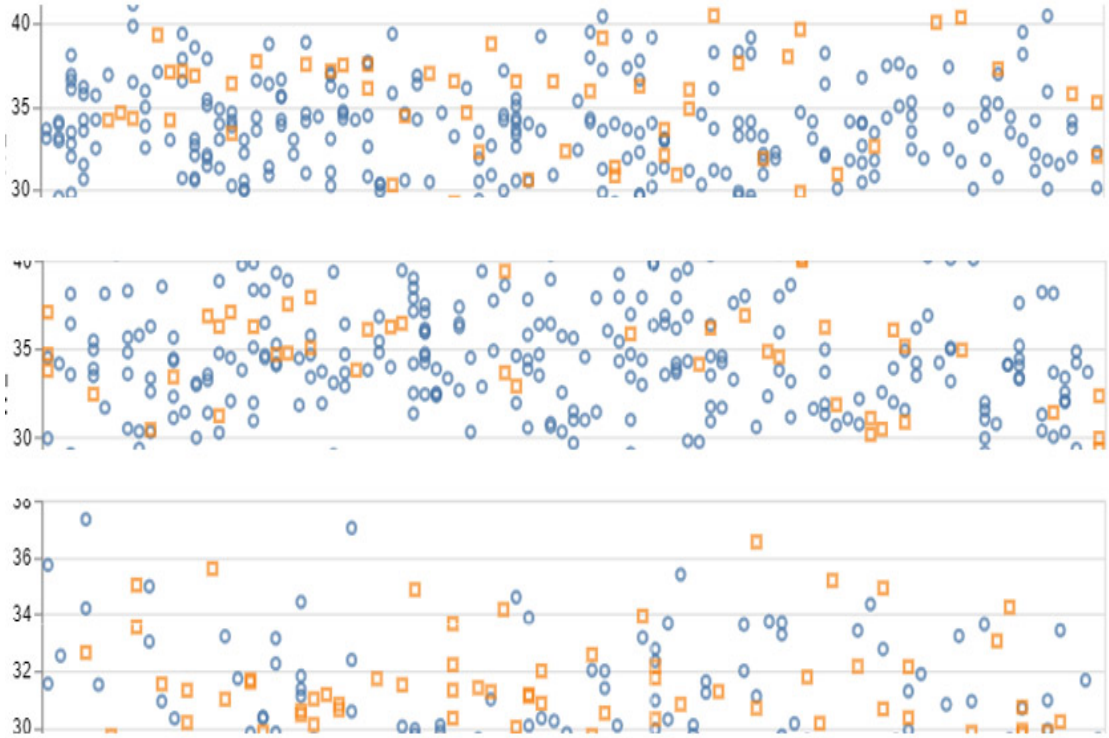

Abbildung 4.1: Illustration des Vergleichs der Aufschlagstatistik der Gegner von Nadal (oben), Djokovic (mitte) und Martin (unten) aus den All-Time-Modellen, im Bereich 30-40 Punkte (Nadal und Djokovic) und im Bereich 30-38 (Martin) im Durchschnitt pro Spiel. Die unterschiedliche Skallierung resultiert aus der automatischen Anpassung an die Daten von TFJS-VIS.

immer wieder Matches wegen solcher kurzfristigen Problemen verloren wurden. Diese Änderung in den Matches könnte das das Modell erkannt haben. Ein plausibler Grund für den Anstieg in der richtigen Vorhersage bezüglich immer älterer Gegner konnte nicht gefunden werden und somit würde für die Aufklärung weitere Forschung benötigt.

4.2 Anwendungsgrenzen

Die Anwendung beschränkt sich auf den Vergleich der verschiedenen Modelle, die sich durch ihre Input-Parameter unterscheiden. Somit ist der Fokus nicht darauf gelegt, die bestmögliche Vorhersage zu treffen. Dieses Ziel könnte jedoch mit den Ergebnissen dieser Arbeit verfolgt werden, indem man verschiedene Parameter der einzelnen Modelle verbindet. Darüber hinaus ist diese Arbeit sehr stark von der Datenbasis abhängig, da auf ihr sowohl die Trainingsdaten als auch die Testdaten beruhen. Somit ist es nicht möglich Fehler und etwaige Ungenauigkeiten, die in der Datenbasis vorhanden sein könnten, aus dem Endprodukt der Arbeit zu entfernen. Beispielsweise wurden einige fehlende Werte,

mit dem Wert 0 aufgefüllt, das sollte zwar in dem Umfang von über 200.000 Datensätzen nicht ausschlaggebend sein, trotzdem kann es die eine oder andere Vorhersage verändern.

Der Datensatz ist auf Matches mit dem Format "Best-Of-Three" limitiert, das hauptsächlich auf der ATP-Tour gespielt wird. Jedoch gibt es auch Turniere, wie beispielsweise Grand Slams (Wimbledon, US Open, French Open, Australian Open) und auch den Davis-Cup, die alle während der relevanten Zeit von 2011 bis 2020 im Format "Best-Of-Five" durchgeführt worden sind. Diese Matches waren ebenfalls im Datensatz vorhanden, jedoch auf die ersten 3 Sätze gekürzt, somit ist die Aussagekraft dieses Teils des Datensatzes über "Best-Of-Three" Matches ein wenig gemindert. Trotzdem richtet sich die Applikation durch diese Kürzung nur an Best-Of-Three Matches. Das bedeutet, eine erfolgreiche Vorhersage für "Best-Of-Five" Matches ist mit dieser Applikation nicht möglich.

Die Analyse in dieser Arbeit war auf drei Spieler limitiert. Es liegt nahe, dass die Ergebnisse bei Spielern in derselben Spielklasse ähnlich ausfallen, so wie es bei Djokovic und Nadal der Fall war. Ansonsten sei noch erwähnt, dass die Modelle nur Daten der ATP verwenden. Durch die Unterschiede in den Vorhersagen zwischen Damen-Tennis und Herren-Tennis, die sogar zu unterschiedlichen Quoten führen [BCM21], sind sie wahrscheinlich nicht geeignet, um Matches aus der WTA Tour klassifizieren zu können. Zusätzlich beschränken sich die Daten auf Spieler, die in der Zeit zwischen 2011 und 2020 aktive professionelle Tennisspieler waren. Würden nun Matches mit anderen Spielern abgefragt, die nicht in der Datenbasis vorkommen, wäre es denkbar, dass diese Vorhersagen ungenauer sind.

4.3 Zukünftige Forschung

4.3.1 Möglichkeiten zur Erweiterung der Modelle

Eine interessante Folgeanwendung wäre die Anwendung der Modelle auf Spiele aus der WTA Serie. Ebenfalls interessant wäre eine Analyse der relevantesten Variablen der einzelnen Modelle, beispielsweise durch eine Visualisierung der Funktionsweise und dem Einfluss der einzelnen Variablen auf die Leistung des jeweiligen Modells [MBR20]. Die Resultate dieser Analyse könnten möglicherweise in weiterer Folge dann verwendet werden, um die besten Variablen aus jedem Modell in eines zusammenzufassen und somit die bestmögliche Genauigkeit aus diesem sehr umfangreichen Datensatz zu erzielen. Diese Variablen-Analyse würde möglicherweise auch Antworten zu den offenen Auffälligkeiten der Modelle in dieser Arbeit geben.

Auch denkbar wäre eine Erweiterung dieser Arbeit um einen vergleichbaren weiteren Machine-Learning-Ansatz, der schon Erfolge in dem Bereich der Tennis-Match-Vorhersage hat, wie beispielsweise die logistische Regression [CSW17] oder auch eine Multi-Target Regression [BGB21]. Dadurch könnte ersichtlich werden, wie sehr sich die Kombination aus diesem Datensatz als Trainingsdaten und den neuronalen Netzen als Machine-Learning-Variante für diese Aufgabe im Vergleich eignet.

4.3.2 Visualisierungsmöglichkeiten auf Basis dieser Arbeit

In einer zukünftigen Arbeit könnten noch weitere Spieler hinzugefügt werden. Auch eine groß angelegte Studie über alle Top 100 Spieler der ATP Weltrangliste durchzuführen, wäre denkbar. Das wäre möglich, indem man über die ganzen Daten iteriert und jeweils die zusammengehörigen Daten eines Spielers in einer farbeinheitlichen Form in ein großes Streudiagramm einfügt. Damit kann getestet werden, wie unterschiedlich die Leistung zwischen den Spielern dann wirklich ist. Ebenfalls vorstellbar wäre ein multidimensionales Koordinatensystem durch parallele Koordinaten, wie in Abbildung 4.2 beispielhaft ersichtlich. Bei dieser Variante werden pro Datensatz die einzelnen Variablen als Y-Werte auf den vielen vertikalen Achsen aufgetragen und dann durch Linien dazwischen verbunden, was Trends und Auffälligkeiten in einzelnen Datensätzen sehr gut darstellen kann. Dadurch können Unterschiede in den Genauigkeiten der Modelle in Bezug auf die verschiedenen Spieler (und vielleicht auch Spielerinnen) visualisiert werden [ID90]. Ebenfalls wäre diese Darstellungsform denkbar für die Analyse eines Modells, das die besten Variablen der vier Ausgangsmodelle vereint. Würde man in dieser Darstellung nun die vier Ausgangsmodelle mit dem Idealmodell vergleichen, so würde man mutmaßlich gut sehen, welche Variablen in welchem Zusammenhang stehen und wie dieser Zusammenhang die Genauigkeit des Modells beeinflusst.

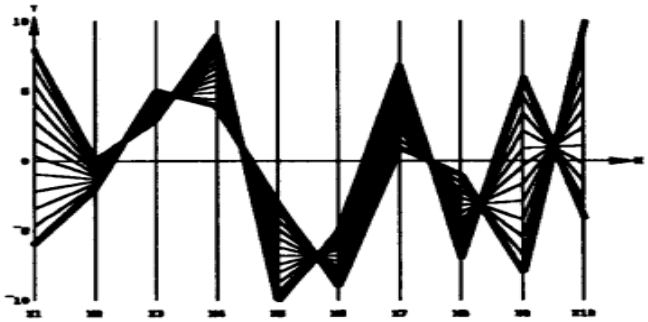

Abbildung 4.2: Beispiel für Visualisierung durch parallele Koordinaten für mehrdimensionale Daten. Bildquelle: [ID90]

Auch von Relevanz für eine sportwissenschaftliche Perspektive wäre eine Visualisierung und Gegenüberstellung der verschiedenen Variablen als Y-Werte, vielleicht nicht als Streudiagramm sondern als Liniendiagramm, wobei jeweils eine Linie alle Matches repräsentiert, mit dem Y-Wert der jeweiligen Variablen. Jede dieser Linien könnte mit einer Farbe versehen werden und so könnten möglicherweise Trends bei bestimmten Parametern entdeckt werden. Diese Ergebnisse könnten Trainer/innen und Spieler/innen im Training weiterhelfen, da zielgenauer trainiert werden kann, wofür Machine Learning zu einem gewissen Grad auch bereits eingesetzt wird [WR17]. Aber auch Wettanbieter/innen und Buchmacher/innen könnten diese Informationen helfen, um genauere Vorhersagen treffen zu können. Somit wäre es möglich, fairere Quoten für die Wettenden anzubieten, da eine geringere Chance besteht, dass Fehler in der Quotenberechnung passieren [HŠŽ19].

Generell wäre es möglicherweise sinnvoll mit den resultierenden Modellen, ähnlich wie Spann et al. [SS09], einen Vergleich durchzuführen. Hierbei könnten die Resultate mit Quoten der Buchmacher/innen und auch anderen Wettmöglichkeiten, wie beispielsweise PM [FVN10], verglichen werden. Zusätzlich wäre denkbar, dass ebenfalls oder alternativ verschiedene, andere Arten von neuronalen Netzwerken in diese Studie miteingebunden werden und somit ihre Leistung auch mitverglichen werden kann. Beispielsweise LSTM Neuronale-Netze oder auch Funktionsanpassende Neuronale-Netze könnten in diesem Kontext ausprobiert werden. Diese Herangehensweisen waren bei aktuellen Vorhersagen und Modellierungen von Covid-19-betreffenden Themen erfolgreich, somit wäre es nur naheliegend, ihre Leistung auch in weiteren Einsatzgebieten zu testen [KCS+21]. Außerdem waren sowohl Feedforward-Neuronale-Netze, als auch generalisierte und auch probabilistische Neuronale-Netze in der Vorhersage von Basketball-Spielen erfolgreich, was zu einem gewissen Grad auch auf mögliche Erfolgschancen bei der Tennis-Match-Vorhersage hindeuten könnte, womit diese drei Arten ebenfalls eine legitime Erweiterung darstellen könnten [LBB09]. Beispielsweise könnten wieder die parallelen Koordinaten genutzt werden, um die Genauigkeiten der verschiedenen Netze darstellen zu können und somit auf den ersten Blick erkennbar zu machen, welche Variante nun am besten abschneidet.

Abbildungen

1.1 Typische Architektur eines neuronalen Netzes 4
1.2 Aufschlagplatzierung nach ML 8

2.1 Unterschiede Genauigkeiten Sportarten 13
2.2 Siegwahrscheinlichkeit von zwei ATP Matches 17
2.3 Beliebtheit visueller Darstellungsmethoden 24
2.4 Visualisierung der Entscheidungsgrenzen 25

3.1 Letzte Epochen des Trainings 33
3.2 Gesamtansicht der Applikation 36
3.3 Streudiagramm Andrej Martin 38
3.4 Auswahl Datum im Streudiagramm 39
3.5 Zusammenfassung der Genauigkeit der Modelle Djokovic 40
3.6 Zusammenfassung der Genauigkeit der Modelle Nadal 42
3.7 Zusammenfassung der Genauigkeit der Modelle Martin 43
3.8 All-Time-Modell und Last-5-Matches-Modell Diagramme von Djokovic. . 44
3.9 All-Time-Modell und Last-5-Matches-Modell Diagramme von Nadal. . . . 45
3.10 Streudiagramme der Modelle Last-Match-Modell und Current-Match-Modell von Nadal 46
3.11 Streudiagramme der Modelle All-Time-Modell und Last-5-Matches-Modell von Martin 47

4.1 Vergleich der Aufschlagleistungen der Gegner 53
4.2 Beispiel für Visualisierung durch Parallel-Koordinaten 55

Tabellen

3.1 Variablen des Datensatzes für jeweils das vorherige Match (), die ganze Karriere (_A) und über die letzten fünf Matches (_L5), diese Daten gibt es jeweils für beide Spieler (_1 und _2) [Nik20] 28
3.2 Variablen des Datensatzes für jeweils das aktuelle Match (_CUR). Diese Daten gibt es jeweils für beide Spieler (_1 und _2) [Nik20] 29
3.3 Die genauen Daten der Modelle im Bezug auf die Vorhersage der Matches von Novak Djokovic 40
3.4 Die genauen Daten der Modelle im Bezug auf die Vorhersage der Matches von Rafael Nadal 41
3.5 Die genauen Daten der Modelle im Bezug auf die Vorhersage der Matches von Andrej Martin 42

Bibliography

[AJO+18] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-Of-The-Art in Artificial Neural Network Applications: A Survey. *Heliyon*, 4(11):e00938, 2018.

[BC05] Tristan Barnett and Stephen R Clarke. Combining Player Statistics to Predict Outcomes of Tennis Matches. *IMA Journal of Management Mathematics*, 16(2):113–120, 2005.

[BCM21] Joxe Maria Barrutiabengoa, Pilar Corredor, and Luis Muga. Does the Betting Industry Price Gender? Evidence from Professional Tennis. *Journal of Sports Economics*, pages 1–37, 2021.

[BDIG15] Jason P Berkowitz, Craig A Depken II, and John M Gandar. Information and Accuracy in Pricing: Evidence from the NCAA Men's Basketball Betting Market. *Journal of Financial Markets*, 25:16–32, 2015.

[BGB21] Firas Bayram, Davide Garbarino, and Annalisa Barla. Predicting Tennis Match Outcomes with Network Analysis and Machine Learning. In *Proceedings of the International Conference on Current Trends in Theory and Practice of Informatics*, pages 505–518. Springer, 2021.

[BK97] Chris Burge and Samuel Karlin. Prediction of Complete Gene Structures in Human Genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, 1997.

[BM01] Indranil Bose and Radha K Mahapatra. Business Data Mining—a Machine Learning Perspective. *Information & management*, 39(3):211–225, 2001.

[BÖ21] Serkan Ballı and Engin Özdemir. A Novel Method for Prediction of Euroleague Game Results using Hybrid Feature Extraction and Machine Learning Techniques. *Chaos, Solitons & Fractals*, 150:111119, 2021.

[Bro12] Alasdair Brown. Evidence of In-Play Insider Trading on a UK Betting Exchange. *Applied Economics*, 44(9):1169–1175, 2012.

[BTP+21] Francesca Bottino, Emanuela Tagliente, Luca Pasquini, Alberto Di Napoli, Martina Lucignani, Lorenzo Figà-Talamanca, and Antonio Napolitano. COVID Mortality Prediction with Machine Learning Methods: a Systematic Review and Critical Appraisal. *Journal of Personalized Medicine*, 11(9):893, 2021.

[BU19] Günal Bilek and Efehan Ulas. Predicting Match Outcome according to the Quality of Opponent in the English Premier League using Situational Variables and Team Performance Indicators. *International Journal of Performance Analysis in Sport*, 10 2019.

[Cac22] Thiem vs. Cachin Quoten, 2022. `https://oddspedia.com/tennis/dominic-thiem-pedro-cachin-935488#odds`, zuletzt zugegriffen am 19/06/2022.

[Cao12] Chenjie Cao. *Sports Data Mining Technology used in Basketball Outcome Prediction.* Masters Thesis. Technological University Dublin, 2012.

[CGC01] Lee Colquitt, Norman H Godwin, and Steven B Caudill. Testing Efficiency across Markets: Evidence from the NCAA Basketball Betting Market. *Journal of Business Finance & Accounting*, 28(1-2):231–248, 2001.

[CMJ+20] Angelos Chatzimparmpas, Rafael Messias Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. In *Computer Graphics Forum*, volume 39, pages 713–756. Wiley Online Library, 2020.

[CMWOA16] Richard G Cowden, Anna Meyer-Weitz, and Kwaku Oppong Asante. Mental Toughness in Competitive Tennis: Relationships with Resilience and Stress. *Frontiers in Psychology*, 7:320, 2016.

[Cow16] Richard G Cowden. Competitive Performance Correlates of Mental Toughness in Tennis: A Preliminary Analysis. *Perceptual and Motor Skills*, 123(1):341–360, 2016.

[CSW17] Andre Cornman, Grant Spellman, and Daniel Wright. Machine Learning for Professional Tennis Match Prediction and Betting. *Project, Stanford Univerisity*, 2017.

[Dat22] H5 Dateierweiterung, 2022. `https://datei.wiki/extension/h5`, zuletzt zugegriffen am 19/06/2022.

[EMS19] Radwa Elshawi, Mohamed Maher, and Sherif Sakr. Automated Machine Learning: State-Of-The-Art and Open Challenges. *arXiv preprint arXiv:1906.02287*, 2019.

[ENM15] Issam El Naqa and Martin J. Murphy. *What Is Machine Learning?*, pages 3–11. Springer International Publishing, Chambridge, 2015.

[FH22] Kai Fischer and Justus Haucap. Home Advantage in Professional Soccer and Betting Market Efficiency: The Role of Spectator Crowds. *Kyklos*, 75(2):294–316, 2022.

[FHS18] Arne Feddersen, Brad R Humphreys, and Brian P Soebbing. Sentiment Bias in National Basketball Association Betting. *Journal of Sports Economics*, 19(4):455–472, 2018.

[FM07] David Forrest and Ian McHale. Anyone for Tennis (Betting)? *The European Journal of Finance*, 13(8):751–768, 2007.

[FVN10] Egon Franck, Erwin Verbeek, and Stephan Nüesch. Prediction Accuracy of Different Market Structures—Bookmakers versus a Betting Exchange. *International Journal of Forecasting*, 26(3):448–459, 2010.

[GBVB21] Youri Geurkink, Jan Boone, Steven Verstockt, and Jan G Bourgois. Machine Learning-based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer. *Applied Sciences*, 11(5):2378, 2021.

[GDBZ98] John M Gandar, William H Dare, Craig R Brown, and Richard A Zuber. Informed Traders and Price Variations in the Betting Market for Professional Basketball Games. *The Journal of Finance*, 53(1):385–401, 1998.

[GK19] Zijian Gao and Amanda Kowalczyk. Random Forest Model identifies Serve Strength as a Key Predictor of Tennis Match Outcome. *Journal of Sports Analytics*, (Preprint):1–8, 2019.

[GLTS09] Eric Gillet, David Leroy, Régis Thouvarecq, and Jean-François Stein. A Notational Analysis of Elite Tennis Serve and Serve-Return Strategies on Slow Surface. *The Journal of Strength & Conditioning Research*, 23(2):532–539, 2009.

[Gol16] Peter Goldsborough. A Tour of Tensorflow. *arXiv preprint arXiv:1610.01178*, 2016.

[GPKR21] Brandon Giles, Peter Peeling, Stephanie Kovalchik, and Machar Reid. Differentiating Movement Styles in Professional Tennis: A Machine Learning and Hierarchical Clustering Approach. *European Journal of Sport Science*, pages 1–10, 2021.

[Grä03] Uwe Grässer. *Emotion, Emotionsverarbeitung und sportliche Leistung: Die Bedeutung des Konstrukts, Klarheit über eigene Gefühle—für Leistungsschwankungen im Tennis.* Universität des Saarlandes, 2003.

[GS08] Ian Graham and H Stott. Predicting Bookmaker Odds and Efficiency for UK Football. *Applied Economics*, 40(1):99–109, 2008.

[HJ20] Tomislav Horvat and Josip Job. The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1380, 2020.

[HPW13] Brad R Humphreys, Rodney J Paul, and Andrew P Weinbach. Consumption Benefits and Gambling: Evidence from the NCAA Basketball Betting Market. *Journal of Economic Psychology*, 39:376–386, 2013.

[HŠŽ19] Ondřej Hubáček, Gustav Šourek, and Filip Železnỳ. Exploiting Sports-Betting Market using Machine Learning. *International Journal of Forecasting*, 35(2):783–796, 2019.

[ID90] Alfred Inselberg and Bernard Dimsdale. Parallel Coordinates: a Tool for Visualizing Multi-Dimensional Geometry. In *Proceedings of the First IEEE Conference on Visualization: (VIS '90)*, pages 361–378. IEEE, 1990.

[JM97] David Jackson and Krzysztof Mosurski. Heavy Defeats in Tennis: Psychological Momentum or Random Effect? *Chance*, 10(2):27–34, 1997.

[JQu21] Jquery-csv, 2021. `https://www.npmjs.com/package/jquery-csv`, zuletzt zugegriffen am 19/06/2022.

[JTS21] Sarika Jain, Ekansh Tiwari, and Prasanjit Sardar. Soccer Result Prediction Using Deep Learning and Neural Networks. In *Intelligent Data Communication Technologies and Internet of Things*, pages 697–707. Springer, 2021.

[KCS+21] Firuz Kamalov, Aswani Cherukuri, Hana Sulieman, Fadi Thabtah, and Akbar Hossain. Machine Learning Applications for COVID-19: A State-Of-The-Art Review. *arXiv preprint arXiv:2101.07824*, 2021.

[KEE+15] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine Learning Applications in Cancer Prognosis and Prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[Ker22] Sequential Model, 2022. `https://keras.io/guides/sequential_model/`, zuletzt zugegriffen am 29/06/2022.

[KPN16] Josua Krause, Adam Perer, and Kenney Ng. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697, 2016.

[KT20] Florian Königstorfer and Stefan Thalmann. Applications of Artificial Intelligence in Commercial Banks–A Research Agenda for Behavioral Finance. *Journal of Behavioral and Experimental Finance*, 27:100352, 2020.

[Lah14] Jiří Lahvička. What causes the Favourite-Longshot Bias? Further Evidence from Tennis. *Applied Economics Letters*, 21(2):90–92, 2014.

[LBB09] Bernard Loeffelholz, Earl Bednar, and Kenneth W Bauer. Predicting NBA Games using Neural Networks. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[LBM19] Sven Lerner, Dipika Badri, and Kevin Monogue. *DeepTennis: Mid-Match Tennis Predictions. In Proceedings of CS230: Deep Learning (Posters).* Stanford University, CA, USA, 2019.

[LF16] Štefan Lyócsa and Igor Fedorko. What drives Intermediation Costs? A Case of Tennis Betting Market. *Applied Economics*, 48(22):2037–2053, 2016.

[Liu20] Xinyu Liu. *Tennis Stroke Recognition: Stroke Classification using Inertial Measuring Unit and Machine Learning Algorithm in Tennis.* Masters Thesis, Delft University of Technology, 2020.

[LN15] Maxwell W Libbrecht and William Stafford Noble. Machine Learning Applications in Genetics and Genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[LV18] Štefan Lyócsa and Tomáš Vỳrost. To Bet or not to Bet: A Reality Check for Tennis Betting Market Efficiency. *Applied Economics*, 50(20):2251–2272, 2018.

[LZH09] Christoph Leitner, Achim Zeileis, and Kurt Hornik. Is Federer Stronger in a Tournament without Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009. *Austrian Journal of Statistics*, 38(4):277–286, 2009.

[MadHAP14] Frank Meyrahn, Iris an der Heiden, Gerd Ahlert, and Holger Preuß. *Wirtschaftsfaktor Sportwetten - Sportfaktor Lotterien. Aktuelle Daten zur Sportwirtschaft.* Bundesministerium für Wirtschaft und Energie, 2014.

[MBR20] Kevin M Mendez, David I Broadhurst, and Stacey N Reinke. Migrating from Partial Least Squares Discriminant Analysis to Artificial Neural Networks: A Comparison of Functionally Equivalent Visualisation and Feature Contribution Tools using Jupyter Notebooks. *Metabolomics*, 16(2):1–15, 2020.

[MC20] Sridharan Manikandan and M Chinnadurai. Evaluation of Students' Performance in Educational Sciences and Prediction of Future Development using Tensorflow. *International Journal of Engineering Education*, 36(6):1783–1790, 2020.

[MD01] Eric Mjolsness and Dennis DeCoste. Machine Learning for Science: State of the Art and Future Prospects. *Science*, 293(5537):2051–2055, 2001.

[MIS+21] Naeem Ahmed Mahoto, Rabia Iftikhar, Asadullah Shaikh, Yousef Asiri, Abdullah Alghamdi, and Khairan Rajab. An Intelligent Business Model for Product Price Prediction Using Machine Learning Approach. *Intelligent Automation & Soft Computing*, 29(3):147–159, 2021.

[MKL+00] Carolyn Mair, Gada Kadoda, Martin Lefley, Keith Phalp, Chris Schofield, Martin Shepperd, and Steve Webster. An Investigation of Machine Learning Based Prediction Systems. *Journal of Systems and Software*, 53(1):23–29, 2000.

[MPM+21] Vishal Meshram, Kailas Patil, Vidula Meshram, Dinesh Hanchate, and S D Ramkteke. Machine Learning in Agriculture Domain: A State-Of-Art Survey. *Artificial Intelligence in the Life Sciences*, 1:100010, 2021.

[Nik20] Nikolay Kashavkin. Kaggle Tennis 2011-2019, 01 2020. `https://www.kaggle.com/datasets/hwaitt/tennis-20112019?select=atp.csv`, zuletzt zugegriffen am 18/06/2022.

[PŠF10] Andrej Panjan, Nejc Šarabon, and Aleš Filipčič. Prediction of the Successfulness of Tennis Players with Machine Learning Methods. *Kinesiology*, 42:98–106, 2010.

[QAHA21] Giorgio Quer, Ramy Arnaout, Michael Henne, and Rima Arnaout. Machine Learning and the Future of Cardiovascular Care. *Journal of the American College of Cardiology*, 77(3):300–313, 2021.

[RGPOM+22] Markel Rico-González, José Pino-Ortega, Amaia Méndez, Filipe Clemente, and Arnold Baca. Machine learning application in soccer: A systematic review. *Biology of Sport*, 40(1):249–263, 2022.

[ROD21] Chris Richter, Martin O'Reilly, and Eamonn Delahunt. Machine Learning in Sports Science: Challenges and Opportunities. *Sports Biomechanics*, pages 1–7, 2021. PMID: 33874846.

[RP22] Arisoa S Randrianasolo and Larry D Pyeatt. Comparing Different Data Representations and Machine Learning Models to Predict Tennis. In *Future of Information and Communication Conference*, pages 488–500. Springer, 2022.

[RRV+20] Nikki Rommers, Roland Rössler, Evert Verhagen, Florian Vandecasteele, Steven Verstockt, Roel Vaeyens, Matthieu Lenoir, Eva D'Hondt, and Erik Witvrouw. A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Medicine and Science in Sports and Exercise*, 52:1745–1751, 2020.

[RSG16] Ryan M Rodenberg, Jeff Sackmann, and Chris Groer. Tennis Integrity: a Sports Law Analytics Review. *The International Sports Law Journal*, 16(1):67–81, 2016.

[SHH19] Alexander Schulz, Fabian Hinder, and Barbara Hammer. Deepview: Visualizing Classification Boundaries of Deep Neural Networks as Scatter Plots using Discriminative Dimensionality Reduction. *arXiv preprint arXiv:1909.09154*, 2019.

[sin17] IPL Visualization and Prediction using HBase, author=Singh, Shubhra and Kaur, Parmeet. *Procedia Computer Science*, 122:910–915, 2017.

[SJGC21] Abhinav Sharma, Arpit Jain, Prateek Gupta, and Vinay Chowdary. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, 9:4843–4873, 2021.

[SM13] Heidrun Schumann and Wolfgang Müller. *Visualisierung: Grundlagen und allgemeine Methoden*. Springer-Verlag, 2013.

[SS09] Martin Spann and Bernd Skiera. Sports Forecasting: a Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters. *Journal of Forecasting*, 28(1):55–72, 2009.

[Suz12] Kenji Suzuki. *Pixel-based Machine Learning in Medical Imaging*, volume 2012. Hindawi, 2012.

[SWXC16] Shaohuai Shi, Qiang Wang, Pengfei Xu, and Xiaowen Chu. Benchmarking State-Of-The-Art Deep Learning Software Tools. In *Proceedings of the 7th International Conference on Cloud Computing and Big Data (CCBD '6)*, pages 99–104. IEEE, 2016.

[TFJ22] TFJS-VIS, 2022. `https://github.com/tensorflow/tfjs/tree/master/tfjs-vis`, zuletzt zugegriffen am 19/06/2022.

[TOK18] Tomoumi Takase, Satoshi Oyama, and Masahito Kurihara. *Effective Neural Network Training with adaptive Learning Rate based on Training Loss*. Volume 101. Elsevier, 2018.

[TROKRP21] Carlos Andrés Tavera Romero, Jesús Hamilton Ortiz, Osamah Ibrahim Khalaf, and Andrea Ríos Prado. Business Intelligence: Business Evolution after Industry 4.0. *Sustainability*, 13(18):10026, 2021.

[TZA19] Fadi Thabtah, Li Zhang, and Neda Abdelhamid. NBA Game Result Prediction using Feature Analysis and Machine Learning. *Annals of Data Science*, 6(1):103–116, 2019.

[VCSVH08] Ben Van Calster, Tim Smits, and Sabine Van Huffel. The Curse of Scoreless Draws in Soccer: The Relationship with a Team's Offensive, Defensive, and Overall Performance. *Journal of Quantitative Analysis in Sports*, 4(1), 2008.

[VW05] Jarke J Van Wijk. The Value of Visualization. In *VIS 05. IEEE Visualization, 2005*, pages 79–86. IEEE, 2005.

[Web22] Webstorm, 2022. `https://www.jetbrains.com/de-de/webstorm/`, zuletzt zugegriffen am 10/07/2022.

[WR17] David Whiteside and Machar Reid. Spatial Characteristics of Professional Tennis Serves with Implications for Serving Aces: A Machine Learning Approach. *Journal of Sports Sciences*, 35(7):648–654, 2017.

[XUD+19] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

[YCHH22] Jack C Yue, Elizabeth P Chou, Ming-Hui Hsieh, and Li-Chen Hsiao. A Study of Forecasting Tennis Matches via the Glicko Model. *Plos One*, 17(4):e0266838, 2022.

[YKJ05] Paul D Yoo, Maria H Kim, and Tony Jan. Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation. In *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC '06)*. Volume 2, pages 835–841. IEEE, 2005.