Aus dem Institut für Diagnostische und Interventionelle Radiologie
der Heinrich-Heine-Universität Düsseldorf
Institutsleiter: Prof. Dr. med. Gerald Antoch

Computerassistierte Differenzierung zwischen Patienten mit idiopathischem Parkinson-Syndrom und gesunden Kontrollprobanden anhand unterschiedlicher MRT-Bildgebungsmodalitäten unter Verwendung von Deep Learning

Dissertation

zur Erlangung des Grades eines Doktors der Medizin der Medizinischen
Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Christian Boschenriedter
2024

Als Inauguraldissertation gedruckt mit Genehmigung der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:
Dekan: Prof. Dr. med. Nikolaj Klöcker
Erstgutachter: Prof. Dr. med. Dipl.-Inform. Julian Caspers
Zweitgutachter: Prof. Dr. med. Simon B. Eickhoff

Zusammenfassung: Obwohl das idiopathische Parkinson-Syndrom eine weit verbreitete und gesundheitsökonomisch relevante Erkrankung ist, besteht ein Mangel an objektiven Biomarkern für die Diagnostik, sodass Fehldiagnosen häufig sind. Da die Bildgebung mittels Magnetresonanztomographie bereits erfolgreich als experimentelle, nichtinvasive Methode zur computerassistierten Erkennung betroffener Patienten genutzt wurde, stellt sie einen vielversprechenden Ansatz für die Entwicklung eines solchen Biomarkers dar. In der vorliegenden Studie wurde daher an einer Hauptkohorte von 135 Probanden (63 Patienten mit idiopathischem Parkinson-Syndrom, 72 gesunde Kontrollprobanden) untersucht, inwieweit strukturelle, funktionelle und diffusionsgewichtete Aufnahmen des Gehirns nach spezifischer Vorverarbeitung für die Differenzierung mittels *deep learning*-Verfahren geeignet sind. Des Weiteren wurde analysiert, inwiefern die Kombination mehrerer Bildgebungsmodalitäten sowie der Einbezug von Angaben zum Alter und Geschlecht der Probanden im Rahmen der Krankheitsvorhersage einen Vorteil gegenüber der Verwendung einer einzigen Bildgebungsmodalität bietet.

Zu diesem Zweck wurden strukturelle T1-gewichtete Aufnahmen mit Methoden aus der Voxel-basierten Morphometrie vorverarbeitet, die Differenzierung erfolgte mit Hilfe eines dreidimensionalen *convolutional neural network*. Aus den funktionellen Aufnahmen im Ruhezustand wurden die Zeitserien (*time courses*) 100 unabhängiger Komponenten mittels dualer Regression extrahiert und unter Verwendung eines *long short-term memory*-Modells klassifiziert. Als Grundlage für die Differenzierung der diffusionsgewichteten Aufnahmen mit einem dreidimensionalen *convolutional neural network* dienten die fraktionale Anisotropie, die radiale Diffusionsfähigkeit sowie die axiale und die mittlere Diffusivität. Die Klassifizierungsergebnisse der drei Bildgebungsmodalitäten wurden in einem multimodalen Ansatz mit probandenspezifischen Angaben zum Alter und Geschlecht kombiniert, um durch Anwendung von *logistic regression* bzw. *support vector machine* ein Gesamtklassifizierungsergebnis zu berechnen. Evaluiert wurden alle Verfahren an der Hauptkohorte mit *hold-out validation* im *80:20 split* und 5-facher Kreuzvalidierung. Deren Generalisierbarkeit wurde im Rahmen einer externen Validierung an mehreren unabhängigen Kohorten überprüft.

Es zeigte sich, dass der multimodale Ansatz in Kombination mit Angaben zum Alter und Geschlecht der Probanden den untersuchten unimodalen Verfahren überlegen ist (*ROC AUC* von 0,91 und *balanced accuracy* von 85 % in der 5-fachen Kreuzvalidierung). Ein Vergleich der einzelnen Bildgebungsmodalitäten untereinander ergab, dass mit den beschriebenen Verfahren anhand der funktionellen Aufnahmen des Gehirns im Ruhezustand eine präzisere Differenzierung der Probanden möglich war als mit Hilfe der anderen beiden Bildgebungsmodalitäten. Ferner zeigte sich, dass die entwickelten Verfahren im Rahmen der externen Validierung eine vergleichsweise geringe Vorhersagegenauigkeit aufwiesen.

Die unzureichende externe Validität bei der Verwendung künstlicher Intelligenz in der Radiologie ist ein häufiges Phänomen, das auf hardware- sowie probandenspezifischen Faktoren beruht. Folglich zeigen diese Ergebnisse, dass eine Anwendung der vorgestellten Verfahren im klinischen Alltag verfrüht wäre. Zur Erhöhung der Klassifizierungsgenauigkeit sind weitere Datensätze von Probanden aus unterschiedlichen Bildgebungsstudien sowie eine Optimierung der Vorverarbeitungsschritte und der *deep learning*-Modelle erforderlich.

Abstract: Despite idiopathic Parkinson's disease being a common disorder with a significant health and economic burden, biomarkers for its diagnosis remain elusive, resulting in frequent misdiagnoses. Magnetic resonance imaging techniques have already been effectively employed as an experimental, non-invasive method for computer-assisted detection of affected patients. Therefore, it offers a promising solution for developing a biomarker for diagnosing Parkinson's disease. In this research study, a cohort of 135 subjects (63 patients diagnosed with idiopathic Parkinson's disease, 72 healthy control subjects) was examined to evaluate structural, functional, and diffusion-weighted brain images. After specific preprocessing, these images were prepared for classification using deep learning methods. Furthermore, it was explored whether combining multiple imaging modalities and including information regarding the subject's age and sex offered any advantages in classification compared to using a single imaging modality.

Structural image preprocessing included voxel-based morphometry techniques, and classification was carried out using a three-dimensional convolutional neural network. The functional images were preprocessed by extracting the time courses of 100 independent components using dual regression and were classified using a long short-term memory model. Furthermore, fractional anisotropy and radial, axial, and mean diffusivity for the diffusion-weighted images served as the basis for classification using a three-dimensional convolutional neural network model. Finally, the results of all three imaging modalities were merged with subject-specific information on age and sex to evaluate an overall classification result using logistic regression and support vector machine models. All methods were applied to the cohort with hold-out validation in an 80:20 split and 5-fold cross-validation. Moreover, generalization was assessed by external validation on several geographically external cohorts.

It turned out that the multimodal approach, combined with information on age and sex, outperformed all unimodal methods (ROC AUC of 0.91 and balanced accuracy of 85% in 5-fold cross-validation). Furthermore, upon comparing individual imaging modalities, it was observed that functional brain images attained higher precision in categorizing subjects than the other two imaging modalities. However, while attempting to classify subjects from external cohorts, all developed methods significantly reduced prediction accuracy.

Insufficient external validity is a common phenomenon when using artificial intelligence models in radiology based on hardware- and subject-specific parameters. Consequently, the results demonstrate that applying the presented methods in clinical practice would be premature. Finally, it can be concluded that additional datasets of subjects from various imaging studies and optimization of the preprocessing steps and deep learning models are required to enhance classification accuracy.

Abkürzungsverzeichnis

ADNI Alzheimer's Disease Neuroimaging Initiative

BOLD blood oxygen level dependent

CNN convolutional neural network

DTI diffusion tensor imaging

EPI echo planar imaging

FWHM full width at half maximum

GK gesunder Kontrollproband oder gesunde Kontrollprobanden

HHU Heinrich-Heine-Universität Düsseldorf

IC independent component

ICA independent component analysis

IPS idiopathisches Parkinson-Syndrom

KI künstliche Intelligenz

LSTM long short-term memory

LR logistic regression

MRT Magnetresonanztomographie

NPV negative predictive value, negativer prädiktiver Wert

PPMI Parkinson's Progression Markers Initiative

PPV positive predictive value, positiver prädiktiver Wert

ReLU rectified linear unit

RNN recurrent neural network

ROC AUC area under the receiver operating characteristic curve

rs-fMRT funktionelle Magnetresonanztomographie im Ruhezustand

SVM support vector machine

TE echo time

TR repetition time

Inhaltsverzeichnis

1 Einleitung **1**
- 1.1 Idiopathisches Parkinson-Syndrom . . . 1
- 1.2 Künstliche Intelligenz in der Medizin . . . 3
- 1.3 Ziele der Studie und Vorgehen . . . 4

2 Grundlagen **7**
- 2.1 Deep Learning . . . 7
  - 2.1.1 Multilayer Perceptron . . . 7
  - 2.1.2 Convolutional Neural Networks . . . 8
  - 2.1.3 Recurrent Neural Networks . . . 10
  - 2.1.4 Training . . . 11
  - 2.1.5 Validierung . . . 12
- 2.2 Physik der MRT-Bildgebung . . . 13
  - 2.2.1 Strukturelle MRT-Bildgebung . . . 14
  - 2.2.2 Funktionelle MRT-Bildgebung . . . 15
  - 2.2.3 Diffusionsgewichtete MRT-Bildgebung . . . 16
- 2.3 Methoden der Bildverarbeitung . . . 17
  - 2.3.1 NIfTI-Standard . . . 17
  - 2.3.2 Brain Extraction . . . 18
  - 2.3.3 Bildregistrierung . . . 18
  - 2.3.4 Voxel-basierte Morphometrie . . . 20
  - 2.3.5 Bewegungskorrektur . . . 21
  - 2.3.6 Intensity Normalization . . . 22
  - 2.3.7 Spatial Smoothing . . . 22
  - 2.3.8 Temporal Filtering . . . 23
- 2.4 Funktionelle Konnektivität . . . 23
  - 2.4.1 Unabhängige Komponentenanalyse . . . 24
  - 2.4.2 Duale Regression . . . 25
- 2.5 Studienspezifisches Matching . . . 26
  - 2.5.1 Confounder . . . 27
  - 2.5.2 Matching-Algorithmus . . . 27

3 Material und Methoden **29**
- 3.1 Datensammlung . . . 29
  - 3.1.1 Interne Probandenkohorte . . . 29
  - 3.1.2 Externe Probandenkohorten . . . 31
- 3.2 Uni- und multimodale Modelle auf interner Kohorte . . . 32
  - 3.2.1 T1-Classifier . . . 32
  - 3.2.2 rs-fMRT-Classifier . . . 35
  - 3.2.3 DTI-Classifier . . . 37
  - 3.2.4 Multimodaler Classifier . . . 39
- 3.3 Externe Validierung der Modelle . . . 40
- 3.4 Untersuchung der Generalisierbarkeit auf multizentrische Kohorten . . . 41
  - 3.4.1 T1-Classifier bei idealer Alters- und Geschlechterverteilung . . . 41
  - 3.4.2 rs-fMRT-Classifier mit Leave-One-Site-Out-Ansatz . . . 42

3.5 Modellbewertung 42
3.6 Implementation und verwendete Software 42

4 Ergebnisse **45**
4.1 Uni- und multimodale Modelle auf interner Kohorte 45
4.1.1 T1-Classifier 45
4.1.2 rs-fMRT-Classifier 46
4.1.3 DTI-Classifier 46
4.1.4 Multimodaler Classifier 46
4.2 Externe Validierung der Modelle 50
4.3 Untersuchung der Generalisierbarkeit auf multizentrische Kohorten 51
4.3.1 T1-Classifier bei idealer Alters- und Geschlechterverteilung 52
4.3.2 rs-fMRT-Classifier mit Leave-One-Site-Out-Ansatz 52

5 Diskussion **53**
5.1 Überblick 53
5.2 Vergleich zu anderen Studien 55
5.2.1 Strukturelle Aufnahmen 55
5.2.2 Funktionelle Aufnahmen 57
5.2.3 Diffusionsgewichtete Aufnahmen 58
5.2.4 Multimodale Ansätze 59
5.3 Klinische Anwendung 60
5.3.1 Relevanz 60
5.3.2 Validität 61
5.3.3 Verwendung als Screening-Biomarker 61
5.4 Lösung der Black Box-Problematik 62
5.5 Limitationen 63
5.5.1 Stichprobengröße 63
5.5.2 Site Effects 64
5.5.3 Patientenseitige Confounder 65
5.5.4 Ground Truth 65
5.5.5 Softwaretechnische Herausforderungen 66
5.6 Schlussfolgerungen 67
5.7 Ausblick 67

6 Anhang **69**
6.1 Literatur- und Quellenverzeichnis 69

7 Danksagung

1 Einleitung

1.1 Idiopathisches Parkinson-Syndrom

Das idiopathische Parkinson-Syndrom (IPS) ist nach der Alzheimer-Krankheit die zweithäufigste neurodegenerative Erkrankung und betrifft in den Industriestaaten ca. 1 % der Bevölkerung im Alter von über 60 Jahren (A. Lee und Gilbert, 2016). Es zählt zur Gruppe der Parkinson-Syndrome, die durch das gleichzeitige Vorliegen einer Bradykinesie und eines Ruhetremors oder Rigors definiert sind (Ogawa et al., 2018). Trotz großer Forschungsbemühungen werden zahlreiche Aspekte der Erkrankung nach wie vor nicht ausreichend verstanden, was zur Folge hat, dass ca. 20 bis 35 % der gestellten IPS-Diagnosen inkorrekt sind (Rizzo et al., 2016; Hustad et al., 2018). In der vorliegenden Studie werden daher neuartige Ansätze zur computerassistierten Diagnostik vorgestellt, bei denen unterschiedliche bildgebende Verfahren genutzt werden.

Von 1950 bis 2015 hat sich die Anzahl an Patienten mit IPS weltweit auf über 6 Millionen verdoppelt und bis 2040 wird eine weitere Verdoppelung erwartet. Diese Entwicklung wird u. a. auf eine steigende Lebenserwartung und eine erhöhte Exposition gegenüber Insektiziden sowie anderen industriellen Schadstoffen zurückgeführt (Dorsey et al., 2018). Neben einer erheblichen Beeinträchtigung der Lebensqualität der Betroffenen verursacht IPS auch eine ökonomische Belastung aller Länder weltweit. Kowal et al. (2013) berechneten für die Vereinigten Staaten von Amerika, dass die Gesamtkosten pro Patient mit IPS im Jahr 2010 um 22.800 $ höher lagen als bei einem vergleichbaren Einwohner ohne IPS. Ungefähr 44 % dieses Betrags entfielen auf indirekte Kosten, beispielsweise auf Grund von eingeschränkter Arbeitsfähigkeit. Studien zu Krankheitskosten aus Deutschland sind weniger aktuell, jedoch berechneten Reese et al. (2011) anhand im Jahr 2006 erhobener Daten, dass pro Patient 16.800 Euro Gesamtkosten innerhalb eines Jahres entstanden sind, wovon 31 % indirekte Kosten ausmachten. Somit ist IPS aus gesundheitsökonomischer Sicht eine relevante Erkrankung, für die weiterhin ein hoher Forschungsbedarf im Hinblick auf Krankheitsentstehung, Diagnostik und Therapie besteht.

Das zentrale Symptom bei betroffenen Patienten ist die Bradykinesie, die durch ein erschwertes und verlangsamtes Ablaufen von Willkürbewegungen gekennzeichnet ist. Der Grund dafür liegt in einer Degeneration von Neuronen in der Substantia nigra (Kanda und Uchida, 2014), die den Neurotransmitter Dopamin synthetisieren und ihn axonal zum Putamen transportieren. Von dort aus führen parallel zueinander ein motorikfördernder und ein motorikhemmender Verschaltungsweg zum motorischen Thalamus. Der aus dieser Degeneration resultierende Dopaminmangel bewirkt eine Disfazilitation des motorikfördernden Wegs und eine Enthemmung des motorikhemmenden Wegs, was zur Folge hat, dass thalamokortikale Neurone verstärkt gehemmt werden. Dies wiederum erschwert die Durchführung von Bewegungen (Alexander und Crutcher, 1990). Die Mechanismen, die zur selektiven Zerstörung der dopaminergen Neuronen führen, sind aktuell nicht befriedigend geklärt (Iheagwam und Etefia, 2019). Unterstützende Kriterien zum Ausschluss anderer Parkinson-Syndrome umfassen einen einseitigen Beginn und eine persistierende Asymmetrie im weiteren Krankheitsverlauf, einen klassischen Ruhetremor, positives Ansprechen auf das Medikament Levodopa sowie eine langsame klinische Progression (Zach et al., 2017). Zu den fakultativen Begleitbeschwerden aller Parkinson-Syndrome zählen sensorische Beschwerden, vegetative Anomalien, psychische

Symptome und kognitive Defizite.

Ein weiteres neuropathologisches Kennzeichen der Erkrankung ist neben dem Verlust dopaminerger Neuronen in der Substantia nigra die Ausbildung intrazellulärer Einschlusskörperchen der Nervenzellen, den sog. Lewy-Körperchen (Forster und Lewy, 1912), die überwiegend aus dem Protein α-Synuclein bestehen. Obwohl der Nachweis dieses Merkmals den Goldstandard in der Diagnostik darstellt (Adler et al., 2014), sind die zur Probengewinnung nötigen Maßnahmen auf Grund ihrer Invasivität in vivo nicht durchführbar, sodass die Diagnose des IPS in der Regel anhand der klinischen Symptomatik gestellt wird (Zach et al., 2017). Durch Symptomüberlappungen mit Krankheitsbildern wie dem essentiellen Tremor, der Multisystematrophie (Palma et al., 2018) oder der progressiven supranukleären Blickparese (Quattrone et al., 2018) sind Fehldiagnosen leicht möglich, zudem treten die diagnostisch relevanten Motorikstörungen erst in einem weit fortgeschrittenen Stadium der Neurodegeneration auf. Nichtmotorische Symptome wie Obstipation, Hypotension oder Depressionen manifestieren sich zwar bereits mehrere Jahre vor dem Einsetzen von Motorikstörungen (Schrag et al., 2015), sind jedoch auf Grund ihrer Unspezifität bisher höchstens in unterstützender Funktion für die Diagnosestellung geeignet. Rizzo et al. (2016) stellten entsprechend in einer Metaanalyse fest, dass die Erkrankung nur in 80 % aller Fälle korrekt diagnostiziert wurde und die Diagnosegenauigkeit über einen Zeitraum von 25 Jahren nicht signifikant verbessert werden konnte. Im Jahr 2018 zeigten Hustad et al. (2018), dass bei den Probanden einer Studie die IPS-Diagnose sogar nur bei 65 % aller Fälle korrekt gestellt wurde.

Es besteht folglich ein dringender Bedarf an sicheren Biomarkern für die Diagnostik. Diese sind gemäß einer Studie von Delenclos et al. (2016) auch für eine bessere klinische Versorgung und die Entwicklung neuer Therapien notwendig, da die Diagnosestellung vereinfacht und der Krankheitsverlauf besser überwacht werden kann, wodurch eine stark personalisierte Behandlung möglich wird. Ein denkbarer Ansatzpunkt sind Gensequenzierungen, da Mutationen des Glucocerebrosidase-Gens bereits als mögliche Ursache der Erkrankung identifiziert und bei 5 bis 10 % aller Patienten nachgewiesen wurden (Beavan et al., 2015). Bei der Suche nach biochemischen Biomarkern sind u. a. der Nachweis und die weitere Analyse von α-Synuclein in Körperflüssigkeiten sowie in peripheren Geweben von Forschungsinteresse. Malek et al. (2014) schlussfolgerten dazu in einer systematischen Übersichtsarbeit, dass dieses Protein zwar weder im Liquor cerebrospinalis noch im Blutplasma ein verlässlicher Marker ist, jedoch dessen Nachweis im enterischen und autonomen Nervensystem das Potential hat, als Surrogatmarker für eine Synucleinopathie des Gehirns zu fungieren.

Durch die Weiterentwicklung bildgebender Verfahren konnten neuroanatomische Biomarker entdeckt werden, die heutzutage im klinischen Alltag bereits Anwendung finden. Die Beurteilung des Dopaminsystems ist nach intravenöser Injektion einer radioaktiv markierten Tracersubstanz, beispielsweise im Rahmen einer Dopamintransporter-Szintigraphie oder einer ^{18}F-DOPA-Positronen-Emissions-Tomographie, möglich (McNeill et al., 2013). Der Tracer reichert sich dabei an präsynaptischen Dopamintransportern bzw. DOPA-Decarboxylasen an, kann mittels Gammakamera nachgewiesen werden und ermöglicht somit die quantitative Messung und räumliche Verteilung seiner molekularen Zielstrukturen. Ravina et al. (2012) konnten zeigen, dass sich mittels Dopamintransporter-Bildgebung bei betroffenen Patienten bereits Jahre vor dem Auftreten von Motorikstörungen entsprechende Veränderungen feststellen lassen.

Die Magnetresonanztomographie (MRT) weist eine breitere Verfügbarkeit auf als die Szintigraphie oder die Positronen-Emissions-Tomographie, dient bei der Diagnostik bisher jedoch lediglich dem Ausschluss sekundärer Ursachen von Parkinsonismus (Frederick und Meijer,

2014). Konventionelle MRT-Aufnahmen können dabei unterstützend genutzt werden, um IPS von atypischen Parkinson-Syndromen abzugrenzen, jedoch sind hirnstrukturelle Veränderungen bei Patienten mit IPS im Frühstadium damit kaum nachweisbar (Politis et al., 2017). Für fortgeschrittene Stadien der Erkrankung konnten Summerfield et al. (2005) jedoch eine Reduktion des Kortexvolumens in bestimmten Hirnregionen im Vergleich zu gesunden Kontrollprobanden (GK) feststellen. Ulla et al. (2013) haben in einer Beobachtungsstudie gezeigt, dass der mittels einer bestimmten MRT-Sequenz gemessene Eisengehalt in Substantia nigra und Putamen bei Patienten mit IPS innerhalb von drei Jahren signifikant angestiegen war, während er bei GK unverändert blieb. Zudem entdeckten sie eine positive Korrelation zwischen der gemessenen Veränderung und dem Auftreten von Motorikstörungen. Die Ergebnisse dieser Studien unterstreichen die potentielle Rolle der MRT bei der Diagnostik von IPS sowie bei der Identifizierung von Krankheitsprogression.

1.2 Künstliche Intelligenz in der Medizin

Im Allgemeinen werden unter dem Begriff 'künstliche Intelligenz' (KI) Computersysteme zusammengefasst, die in der Lage sind, nichtphysische Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern (Iqbal und Vinay, 2022). Zu diesem Zweck werten entsprechende Systeme digital vorliegende Informationen aus, um beispielsweise Diagnosen oder Therapien vorzuschlagen. Durch die Verwendung von KI in der Medizin können somit nicht nur Ärzte entlastet, sondern es kann auch die Versorgung individueller Patienten verbessert werden (Buch et al., 2018). Zum einen übertrifft KI bei der Bearbeitung bestimmter, klar definierter Klassifizierungsaufgaben teilweise erfahrene Ärzte (Esteva et al., 2017; Lakhani und Sundaram, 2017) und zum anderen ist diese Technik besser verfügbar bzw. kosteneffizienter als menschliche Experten. Dieser Vorteil ist vor allem in Orten mit geringer Ärztedichte und knappen finanziellen Mitteln von Bedeutung.

Die beachtliche Vorhersagegenauigkeit von KI in der Medizin beruht hauptsächlich auf der Verwendung von *deep learning*, da es mit dieser Technik im Gegensatz zum zuvor vorherrschenden klassischen maschinellen Lernen (*machine learning*) nicht mehr nötig ist, die für eine Vorhersage benötigten Merkmale (*features*) aus den Eingabedaten explizit zu definieren (Hosny et al., 2018). Auf diese Weise konnte die automatisierte Bilderkennung deutlich verbessert werden, was vor allem für eine Anwendung von *deep learning* in der Radiologie spricht. Beispielsweise kann diese Technik dafür eingesetzt werden, um konventionelle Röntgenaufnahmen (Guan et al., 2018; Shen et al., 2019; C. Lee et al., 2020), aber auch dreidimensionale Bilddaten (Jnawali et al., 2018; Polat und Danaei Mehr, 2019) zu klassifizieren. Daneben können KI-Methoden bzw. *deep learning* auch zur Lösung weiterer praktischer Probleme genutzt werden, darunter die Priorisierung zu sichtender Aufnahmen, die Verbesserung der Bildrekonstruktion aus den Rohdaten oder die Erkennung von Müdigkeit bei dem befundenden Radiologen (Thrall et al., 2018).

Obwohl *deep learning* eine junge Forschungsdisziplin in der Medizin ist, gewinnt sie durch die wachsende Verfügbarkeit medizinischer Datenbanken und Rechenleistung sowie durch die daraus resultierende steigende Anzahl an Publikationen fachübergreifend an Popularität (Ching et al., 2018). Bemerkenswerte Resultate erzielten beispielsweise Sahlsten et al. (2019) mit einem Algorithmus, der die diabetische Retinopathie und Makulaödeme anhand von Fundusaufnahmen detektiert. Courtiol et al. (2019) entwickelten ein Modell, mit dem auf Grundlage digitalisierter histopathologischer Schnittbilder von malignen Mesotheliomen die Überlebensrate betroffener Patientengruppen vorhergesagt werden kann. Es ist daher zu vermuten, dass *deep learning* in Zukunft verstärkt in die klinische Diagnostik nahezu aller Fachbereiche Einzug halten und zu einer Entlastung der behandelnden Ärzte und gleichzeitig

zu einer Erhöhung der Diagnosesicherheit führen wird.

Für die KI-gestützte Diagnostik von IPS steht eine Reihe von experimentellen Ansätzen zur Verfügung, die teilweise stark voneinander abweichen und auch nicht zwingend auf Bildgebung basieren, sondern sehr unterschiedliche Eingabedaten verarbeiten können. Beispielsweise verwendeten Pereira et al. (2016) ein auf *deep learning* basierendes Verfahren zur Handschriftanalyse. Dabei mussten die Probanden einfache Figuren mit einem speziellen biometrischen Stift nachzeichnen, der mit unterschiedlichen Sensoren ausgestattet war, um Parameter wie den Auflagedruck auf das Papier oder die Griffstärke messen zu können. Oh et al. (2018) nutzten die Messsignale der Elektroenzephalographie mit 14 Kanälen, um IPS zu detektieren. In einem anderen Ansatz setzten Eskofier et al. (2016) inertiale Messeinheiten zur exakten Bewegungsanalyse ein, während ihre Probanden unterschiedliche motorische Übungen durchführten. Die auf diese Weise erhobenen Daten wurden sowohl mit konventionellem *machine learning* als auch mit *deep learning* ausgewertet, wobei Letztgenanntes in Bezug auf die Klassifikationsgenauigkeit laut den Autoren deutlich überlegen war.

1.3 Ziele der Studie und Vorgehen

In der vorliegenden Studie wurde retrospektiv untersucht, inwieweit eine Differenzierung zwischen Patienten mit IPS und GK anhand einer mehrsequentiellen MRT-Untersuchung des Gehirns mit Hilfe von an die jeweilige Bildgebungsmodalität angepassten *deep learning*-Verfahren möglich ist. Das übergeordnete Ziel besteht darin, einen objektiven Biomarker für die IPS-Diagnostik zu entwickeln, der weder die Verwendung von radioaktiven Substanzen, wie bei der Dopamintransporter-Szintigraphie, noch einen invasiven Eingriff voraussetzt. Die Suche nach einem solchen Biomarker ist u. a. dadurch gerechtfertigt, dass IPS aktuell vor allem anhand der klinischen Symptomatik diagnostiziert wird, was zu einem beachtlichen Anteil an Fehldiagnosen führt (vgl. Unterkapitel 1.1). Zudem existieren bereits zahlreiche erfolgversprechende experimentelle Ansätze, die die MRT-Bildgebung für diagnostische Zwecke nutzen (vgl. Unterkapitel 5.2). Die jeweiligen Autoren berücksichtigten jedoch in der Regel lediglich eine einzige Bildgebungsmodalität, untersuchten nur kleine Probandenkollektive, die keine Ableitung allgemeingültiger Aussagen zulassen, oder verzichteten auf eine Validierung ihrer Verfahren an Aufnahmen aus externen Studien.

Die Klassifizierung der vorverarbeiteten MRT-Aufnahmen erfolgte in der vorliegenden Studie unter Verwendung von *deep learning*, da dieses Verfahren bereits häufig eine überlegene Leistung bei der Analyse medizinischer Bilder offenbarte (Cai et al., 2020) und sich generell dadurch auszeichnet, dass damit aus großen Datenmengen die für eine Klassifizierung relevanten *features* eigenständig extrahieren werden können (Shaheen et al., 2016). Ein weiteres Ziel der vorliegenden Studie bestand darin zu untersuchen, inwieweit sich die Ergebnisse aus der Analyse einer Hauptkohorte bei externen Kohorten reproduzieren lassen bzw. warum sie nur eingeschränkt reproduzierbar sind. Zu diesem Zweck wurde ein multimodaler Untersuchungsansatz verwendet, der je Proband idealerweise jeweils eine strukturelle, eine funktionelle und eine diffusionsgewichtete MRT-Aufnahme zur Klassifizierung (Patient mit IPS oder GK) berücksichtigte. Jede einzelne der genannten Bildgebungsmodalitäten wurde von anderen Autoren in experimentellen Ansätzen bereits erfolgreich für die Erkennung von IPS verwendet (vgl. Chakraborty et al., 2020; Tian et al., 2020; H. Zhao et al., 2022). Außerdem wurde untersucht, ob eine Kombination mehrerer Bildgebungsmodalitäten und Angaben zum Alter und Geschlecht der Probanden im Vergleich zur Verwendung einer einzelnen Bildgebungsmodalität einen Vorteil bei der Klassifizierung mit sich bringt. Schließlich wurde eruiert, ob eine Verwendung dieser Verfahren in der klinischen Diagnostik sinnvoll ist. Die zu klassifizierende Kohorte bestand primär aus Probanden einer Bildgebungsstudie der

Heinrich-Heine-Universität Düsseldorf (HHU). Weitere, davon unabhängige Kohorten wurden im Rahmen einer geographisch externen Validierung eingeschlossen.

Die Zuordnung, ob eine bestimmte Aufnahme von einem Patienten mit IPS oder GK stammt, erfolgte mit Hilfe eigens entwickelter Klassifikatoren ('Classifier'), die vor der eigentlichen Klassifizierung auch eine Vorverarbeitung der Aufnahme durchführten. In einem ersten Ansatz wurde mit einem 'T1-Classifier' (Abschnitt 3.2.1) überprüft, in welchem Umfang eine Klassifizierung anhand struktureller T1-gewichteter MRT-Aufnahmen möglich ist. Dazu wurden Aufnahmen aus der HHU-Kohorte in Trainings- und Testdaten aufgeteilt (*hold-out validation*), der Classifier unter Verwendung der Trainingsdaten entwickelt und schließlich an den Testdaten validiert. Zur Überprüfung der internen Generalisierbarkeit erfolgte eine 5-fache Kreuzvalidierung an der Kohorte. Zur externen Validierung wurde der mit den Aufnahmen aus der HHU-Kohorte trainierte Classifier wiederum an zwei externen Kohorten getestet (Unterkapitel 3.3). Des Weiteren wurde mit den Aufnahmen aus externen Datenbanken ein weiterer Classifier zur Erkennung von Patienten mit IPS trainiert und an der HHU-Kohorte getestet (Abschnitt 3.4.1). Abschließend wurden mittels dreidimensionaler *class activation maps* die Bereiche des Gehirns visualisiert, die der Classifier zur Vorhersage genutzt hatte.

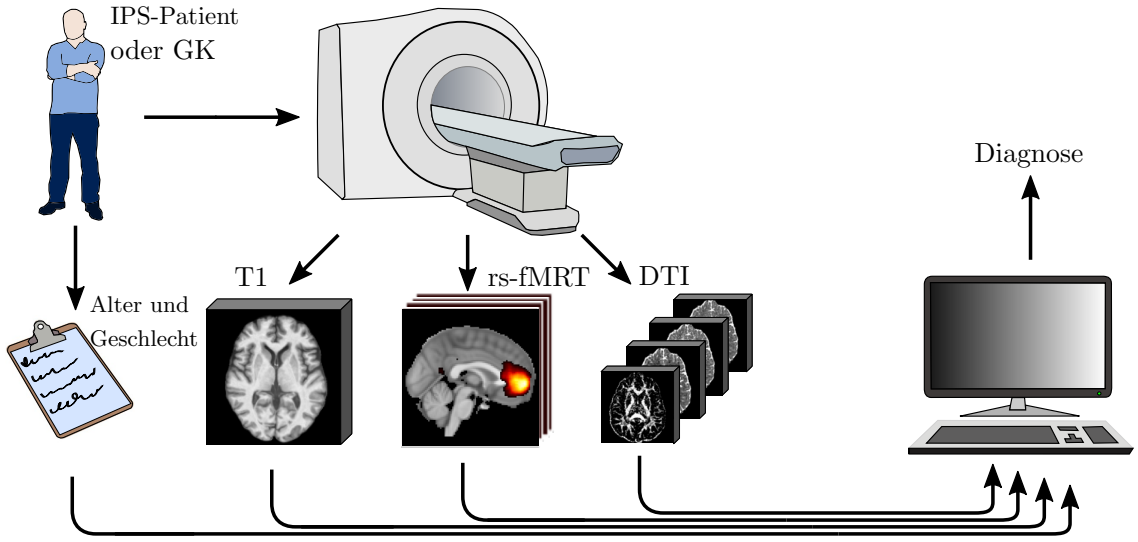

Abb. 1.1: **Leitidee der vorliegenden Studie.** Im Rahmen der IPS-Diagnostik erhält der jeweilige Proband eine MRT-Untersuchung. Die dabei akquirierten Aufnahmen umfassen drei Bildgebungsmodalitäten und werden zusammen mit Angaben zum Alter und Geschlecht an einen Server übermittelt. Anhand dieser Informationen wird die Wahrscheinlichkeit für das Vorliegen von IPS berechnet.

In einem zweiten Ansatz wurde ein neuartiges Verfahren zur Klassifizierung funktioneller MRT-Aufnahmen im Ruhezustand (rs-fMRT) an der HHU-Kohorte angewandt ('rs-fMRT-Classifier', Abschnitt 3.2.2). Diese Aufnahmen wurden zunächst umfangreich vorverarbeitet und anschließend mittels duale Regression die sog. Zeitserien (*time courses*) 100 unabhängiger Komponenten extrahiert. Die Entwicklung und Validierung des Classifiers erfolgten für die HHU-Kohorte analog zum T1-Classifier mittels *hold-out validation* und 5-facher Kreuzvalidierung. Zur externen Validierung wurde der mit rs-fMRT-Aufnahmen der HHU-Kohorte trainierte Classifier an zwei externen Kohorten getestet (Unterkapitel 3.3). Außerdem wurden die Aufnahmen aus der HHU-Kohorte um jene aus vier externen Kohorten erweitert, wobei der Classifier mit jeweils vier Kohorten trainiert und an der fünften Kohorte getestet wurde (Abschnitt 3.4.2).

Die diffusionsgewichteten MRT-Aufnahmen wurden in einem dritten Ansatz analysiert. Mit Hilfe der Diffusions-Tensor-Bildgebung (*diffusion tensor imaging*, *DTI*) wurden die Metriken der fraktionalen Anisotropie, der radialen Diffusionsfähigkeit sowie der axialen und der mittleren Diffusivität berechnet, um die strukturelle Konnektivität in den Gehirnen der Probanden darzustellen. Diese *DTI*-Metriken wurden anschließend zur Bestimmung der Gruppenzugehörigkeit (IPS oder GK) unter Verwendung des dafür entwickelten 'DTI-Classifiers' (Abschnitt 3.2.3) genutzt. Analog zu den ersten beiden Ansätzen wurde die Vorhersagegenauigkeit des DTI-Classifiers mittels *hold-out validation* und 5-facher Kreuzvalidierung evaluiert.

Schließlich wurden die sich teilweise widersprechenden Klassifizierungsergebnisse aller drei unimodalen Classifier zusammen mit probandenspezifischen Angaben zum Alter und Geschlecht kombiniert, um damit die definitive Gruppenzugehörigkeit eines jeden Probanden mit Hilfe eines multimodalen Classifiers (Abschnitt 3.2.4) vorherzusagen (siehe Abbildung 1.1). Für diese Berechnung wurde im Gegensatz zu den vorherigen Classifiern ausschließlich klassisches *machine learning* verwendet. Weiterhin wurde mittels Modellkalibrierung untersucht, zu welchem Grad die vorhergesagte Wahrscheinlichkeit für das Vorliegen von IPS mit dem tatsächlichen Vorliegen der Erkrankung übereinstimmt. Die Validierung innerhalb der HHU-Kohorte erfolgte analog zu den anderen Classifiern mittels *hold-out validation* und 5-facher Kreuzvalidierung. Zusätzlich wurde ein mit den Datensätzen der HHU-Kohorte trainierter Classifier an zwei externen Kohorten getestet (Unterkapitel 3.3).

2 Grundlagen

2.1 Deep Learning

Deep learning stellt einen Teilbereich von KI dar, bei dem mehrschichtige neuronale Netze eingesetzt werden. Darunter sind komplexe Algorithmen zu verstehen, mit deren Hilfe große Datenmengen verarbeitet und Muster sowie Zusammenhänge erkannt werden können. Im Gegensatz zu flachen neuronalen Netzen, die nur aus einer einzigen Schicht bestehen, können *deep learning*-Modelle nahezu beliebig komplexe *features* aus Eingabedaten extrahieren, um sie beispielsweise für Klassifizierungen zu nutzen. Anders als im klassischen *machine learning* ist es nicht erforderlich, diese *features* im Vorfeld manuell zu definieren, da sie vom Algorithmus automatisch ermittelt werden. Die folgenden Erläuterungen zu *deep learning* basieren auf Arbeiten von Ramsundar et al. (2019), Goodfellow et al. (2016) und Chollet (2018).

2.1.1 Multilayer Perceptron

Unter einer Klassifizierung wird die Zuordnung von Eingabedaten zu einer Gruppe verstanden. In der vorliegenden Studie bestanden diese Daten z. B. aus den durch MRT erhobenen Messwerten in Form von Voxel-Intensitäten sowie deren zugehörigen Koordinaten. Sie können mathematisch in ihrer Gesamtheit als dreidimensionale Tensoren verstanden werden, die die drei Raumdimensionen abbilden. Ein unkomplizierter Ansatz, um aus diesen Eingabedaten eine zugehörige Gruppe als Ausgabe zu berechnen, ist die Verwendung eines sog. *multilayer perceptron*. Dieses besteht aus einer Eingabeschicht, um Eingabedaten zu empfangen, einer Ausgabeschicht, die eine Vorhersage bzw. Klassifizierung bezüglich der Eingabedaten ausgibt und dazwischen mindestens einer Schicht, innerhalb der die dazu notwendigen Berechnungen durchgeführt werden. Eine Möglichkeit der Verarbeitung der Eingabedaten ist deren Überführen in die Form eines eindimensionalen Vektors $\mathbf{x}$, indem alle Elemente aus dem mehrdimensionalen Tensor in einer festen Reihenfolge in $\mathbf{x}$ aneinandergereiht werden. Dadurch geht zwar die lokale räumliche Beziehung der Voxel zueinander weitgehend verloren, doch die anschließende Verarbeitung wird vereinfacht. Bei binären Klassifizierungen beschränkt sich die Ausgabe auf $y = 1$ bzw. $y = 0$, was beispielsweise einer 'IPS-Gruppe' bzw. 'GK-Gruppe' entspricht. Zur Klassifizierung von $\mathbf{x}$ ist demnach eine Funktion f gesucht, für die Folgendes gilt:

$$y = f(\mathbf{x})$$

Der Ansatz des überwachten *machine learning* besteht darin, dass ein Computeralgorithmus die Funktion f anhand zahlreicher Beispieldatensätze selbstständig 'erlernt'. Dieser Vorgang wird als Training bezeichnet und die dafür genutzten Datensätze bestehen jeweils aus Eingabedaten in der zuvor beschriebenen Form sowie ihrer zugehörigen Gruppe. Zur konkreten Durchführung des Trainings wird ein Modell erstellt, welches aus mindestens einer allgemein definierten Funktion f und den dazugehörigen Funktionsparametern besteht. Letztere können beliebige Zahlenwerte annehmen und bestimmen somit die konkrete Funktion. Während des Trainings sucht ein Algorithmus nach genau solchen Parametern, für die $f(\mathbf{x})$ für möglichst viele Datensätze so nah wie möglich an der zu $\mathbf{x}$ zugehörigen Gruppe y liegt.

Modelle können grundsätzlich unterschiedlich aufgebaut sein. Eine der einfachsten Formen stellt das lineare Modell dar, das durch

$$y = \mathbf{M}\mathbf{x} + \mathbf{b}$$

definiert ist. $\mathbf{M}$ entspricht dabei einer Matrix und $\mathbf{b}$ einem Vektor bzw. bei binären Klassifizierungen einem Skalar. $\mathbf{M}$ und $\mathbf{b}$ enthalten die durch Training erlernbaren Parameter des Modells. Zur besseren Unterscheidung werden die Parameter in $\mathbf{M}$ als *weights* bezeichnet und in $\mathbf{b}$ als *biases*. Mit diesem linearen Modell können Klassifizierungen vorgenommen werden, falls zwischen den Eingabedaten $\mathbf{x}$ und der zugehörigen Gruppe y ein linearer Zusammenhang besteht. Die ist bei komplexeren Eingabedaten wie etwa Bildern, Sprache oder Text jedoch nicht der Fall.

Zur Umgehung dieser Einschränkung wird auf die lineare Transformation $\mathbf{M}_1\mathbf{x} + \mathbf{b}_1$ eine nichtlineare Funktion φ angewendet und mit dem Ergebnis anschließend eine erneute lineare Transformation durchgeführt:

$$y = \mathbf{M}_2\varphi(\mathbf{M}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$$

φ wird als Aktivierungsfunktion bezeichnet und ermöglicht es dem Modell auf Grund ihrer Nichtlinearität, mehr als nur lineare Zusammenhänge darzustellen. Häufig verwendete Aktivierungsfunktionen umfassen beispielsweise *rectified linear unit (ReLU)* mit $\varphi(x) = \max(0, x)$, den hyperbolischen Tangens mit $\varphi(x) = \tanh(x)$ und die logistische Sigmoidfunktion mit $\varphi(x) = 1/(1 + e^{-x})$.

Der Term $\mathbf{h} = \varphi(\mathbf{M}_1\mathbf{x} + \mathbf{b}_1)$ beschreibt somit eine lineare Transformation mit anschließender nichtlinearer Transformation durch φ. Analog zu einem Abschnitt oder einer Schicht, in der eingegebene Daten verarbeitet werden, wird $\mathbf{h}$ als *hidden layer* bezeichnet, da $\mathbf{h}$ auf Grund seiner Funktion als 'versteckte Schicht' oder 'Zwischenschicht' nicht einem Endergebnis y entspricht, sondern sich weitere Berechnungen an $\mathbf{h}$ anschließen. Um die Mustererkennung zu verbessern, kann die Komplexität des Modells erhöht werden, indem weitere *hidden layers* $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$ hinzugefügt werden. Demnach stellt die Ausgabe einer Schicht $\mathbf{h}_1 = \varphi_1(\mathbf{M}_1\mathbf{x}+\mathbf{b}_1)$ zugleich einen Teil der Eingabe der nächsten Schicht $\mathbf{h}_2 = \varphi_2(\mathbf{M}_2\mathbf{h}_1+\mathbf{b}_2)$ dar. Die Verwendung mehrerer *hidden layers* ist das charakteristische Merkmal von *deep learning* und verleiht dem Modell eine umfangreiche innere Struktur bzw. 'Tiefe'. Wenn in diesem Modell an n Schichten Transformationen durchgeführt werden, so entfallen davon $n - 1$ Schichten auf *hidden layers* – die übrige Schicht ist die Ausgabeschicht. Dieses allgemeine Modell stellt den grundlegenden Aufbau des *multilayer perceptron* dar und lässt sich mit der folgenden Formel beschreiben:

$$y = \mathbf{M}_n\mathbf{h}_{n-1} + \mathbf{b}_n$$

Um sicherzustellen, dass y im Kontext einer Klassifizierung ausschließlich Werte zwischen 0 und 1 annimmt, erfolgt abschließend eine Transformation durch die Anwendung der logistischen Sigmoidfunktion φ_S:

$$y = \varphi_S(\mathbf{M}_n\mathbf{h}_{n-1} + \mathbf{b}_n)$$

2.1.2 Convolutional Neural Networks

Convolutional neural networks (*CNNs*) sind eine Klasse von *deep learning*-Modellen, die vor allem bei der Klassifizierung von Bild- und Audiodaten eine hohe Vorhersagegenauigkeit zeigen (Ciregan et al., 2012; Chandna et al., 2017; Xu et al., 2018). Hinsichtlich ihres Aufbaus ähneln sie einem *multilayer perceptron*, jedoch unterscheiden sich die Verarbeitungsschritte innerhalb der einzelnen Schichten. Bei dem *multilayer perceptron* ist die Berechnung jedes

Ausgabeelements einer Schicht abhängig von jedem Element der vorherigen Schicht. *CNNs* verwenden hingegen sog. *convolutional layers*, bei denen alle Elemente für die Ausgabe an die nächste Schicht nacheinander berechnet werden, und zwar anhand der Elemente eines lokal begrenzten Bereichs der vorherigen Schicht. Damit sind *CNNs* in der Lage, lokale Muster zu erkennen, wie etwa horizontale oder vertikale Linien in zweidimensionalen Bildern. Die dabei verwendete Transformation wird als *convolution* bezeichnet und entspricht mathematisch einer diskreten Faltung. Da die räumliche Anordnung der in den Eingabedaten enthaltenen Elemente für die weitere Verarbeitung im *CNN* von Bedeutung ist, werden diese Elemente nicht mehr durch einen eindimensionalen Vektor $\mathbf{x}$ dargestellt, sondern durch den mehrdimensionalen Tensor $\mathbf{X}$. Zur Veranschaulichung wird hier zunächst angenommen, $\mathbf{X}$ sei ein zweidimensionales Bild mit den Dimensionen $m \times n$, auch wenn die in der vorliegenden Studie verwendeten Modelle höherdimensional sind.

Um auf $\mathbf{X}$ einfache, lokale Muster zu identifizieren, wird eine wesentlich kleinere Filtermatrix $\mathbf{F}$ der Größe $k \times l$ benötigt, welche schrittweise über alle Bereiche von $\mathbf{X}$ gleitet. Dabei wird $\mathbf{F}$ zunächst komponentenweise mit dem unterliegenden Ausschnitt von $\mathbf{X}$ multipliziert und im Anschluss werden diese Produkte summiert. Die so entstehende Ausgabematrix $\mathbf{A}$ wird als *feature map* bezeichnet und enthält alle berechneten Summen (vgl. Abbildung 2.1). Mathematisch ist die *convolution* für ein Element i, j von $\mathbf{A}$ definiert als:

$$\mathbf{A}(i,j) = (\mathbf{F} * \mathbf{X})(i,j) = \sum_{m=1}^{k} \sum_{n=1}^{l} \mathbf{X}(i-m, j-n)\mathbf{F}(m,n)$$

Die Filtermatrix $\mathbf{F}$ muss so konstruiert werden, dass sie bei Anwendung auf die Bereiche von $\mathbf{X}$, die das gesuchte Muster enthalten, möglichst hohe Zahlenwerte produziert. Hohe Zahlenwerte auf der *feature map* $\mathbf{A}$ bedeuten folglich, dass in diesem Bereich das gewünschte Muster erkannt wurde. Diese Muster können theoretisch beliebig komplex sein. Um beispielsweise horizontale bzw. vertikale Kanten zu erkennen, reichen die folgenden Filtermatrizen $\mathbf{F}_{\text{hor}}$ und $\mathbf{F}_{\text{ver}}$ aus (Nisha und Sharma, 2015):

$$\mathbf{F}_{\text{hor}} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}, \mathbf{F}_{\text{ver}} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

Die in $\mathbf{F}$ enthaltenen Parameter sind mit den erlernbaren *weights* $\mathbf{M}$ des *multilayer perceptron* vergleichbar, wobei $\mathbf{F}$ wesentlich kleiner ist als $\mathbf{M}$ und somit den Speicherbedarf des Modells reduziert. Nach Durchführung der linearen *convolution* wird schließlich eine nichtlineare Aktivierungsfunktion φ auf das Ergebnis angewandt. Die dabei entstehende Ausgabe wird an die nächste Schicht weitergegeben.

Eine weitere Besonderheit von *CNNs* sind *pooling layers*, die die vorherige Schicht zusammenfassen und damit deren Größe reduzieren. Ein mögliches Verfahren hierfür ist *max-pooling*, das bei Anwendung auf die *feature map* $\mathbf{A}$ diese in nichtüberlappende Rechtecke unterteilt und den Maximalwert jedes Rechtecks ausgibt (vgl. Abbildung 2.1). Dies ermöglicht es einerseits, den Speicherbedarf des Modells weiter zu verringern, andererseits reduziert es eine eventuelle Überanpassung (*overfitting*, vgl. Abschnitt 2.1.5).

Durch wiederholte Anwendung von *convolutional layers* und *pooling layers* auf die Eingabedaten $\mathbf{X}$ können daraus zunehmend komplexere Muster extrahiert werden. Werden bei der Klassifizierung von Gesichtern auf Bildern beispielsweise von einer ersten *feature map* $\mathbf{A}_1$ lediglich Kanten erkannt, identifiziert die folgende $\mathbf{A}_2$ auf dieser Basis *features* wie Augen oder Ohren. Mit der *feature map* $\mathbf{A}_3$ können schließlich komplexe *features*, wie Gesichter, erkannt werden. Um aus allen *features* ein Klassifizierungsergebnis y zu erhalten, wird zum

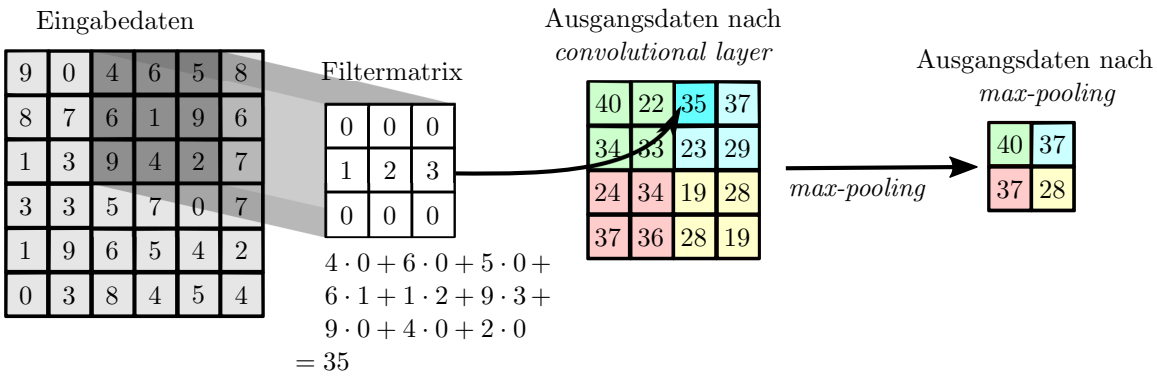

Abb. 2.1: ***Convolution* und *max-pooling*.** Auf die Eingabedaten wird eine 3×3-Filtermatrix angewandt, anschließend erfolgt *max-pooling* mit einem 2×2 großen Filter.

Schluss eine Schicht benötigt, die nicht nur nach lokalen *features* sucht, sondern, ähnlich der Funktionsweise eines *multilayer perceptron*, mit jedem Ausgabeelement der vorherigen Schicht in Verbindung steht. Diese Aufgabe wird durch ein *fully-connected layer* erfüllt. Ist die berechnete Ausgabe mehrdimensional, so muss sie zusätzlich durch sog. *flattening* in einen eindimensionalen Vektor transformiert werden.

Um visuell darstellen zu können, warum ein *CNN* bei der Bilderkennung eine bestimmte Vorhersage getroffen hat, entwickelten Selvaraju et al. (2017) die Methode des *gradient-weighted class activation mapping*. Da die hinteren *layers* des *CNN* zwar am besten in der Lage sind, komplexe Strukturen zu erkennen, auf räumlicher Ebene jedoch ab dem *fully-connected layer* Informationen verloren gehen, verwendeten die Autoren zur weiteren Analyse den hintersten *convolutional layer*. Dieser enthält in Abhängigkeit von der getroffenen Vorhersage die Informationen darüber, welche Regionen im ursprünglichen Bild mit welcher Gewichtung zu der entsprechenden Vorhersage geführt haben. Diese Informationen werden dann in Form einer *heatmap* mit dem ursprünglichen Bild überlagert, sodass die für die Klassifizierung relevanten Bereiche des Bildes durch farbliche Hervorhebungen unmittelbar zu erkennen sind.

2.1.3 Recurrent Neural Networks

Recurrent neural networks (*RNNs*) sind speziell zur Analyse sequenzieller Daten, wie Video- oder Sprachdateien – und somit auch funktioneller MRT-Aufnahmen (vgl. Abschnitt 2.2.2) – geeignet. Sie zeichnen sich dadurch aus, dass *layers* dieser Netzwerke zur Berechnung ihrer Ausgabe $\mathbf{Y}$ auf *layers* bereits vorausgegangener Zeitschritte zurückgreifen können, wodurch die zeitlichen Verläufe der Eingabedaten berücksichtigt werden. $\mathbf{Y}_t$ zum Zeitschritt t setzt sich demnach aus den Eingabedaten $\mathbf{X}_t$ und den Berechnungen eines vorherigen Zeitschritts $\mathbf{Y}_{t-1}$ zusammen:

$$\mathbf{Y}_t = \varphi(\mathbf{X}_t\mathbf{M}_x + \mathbf{Y}_{t-1}\mathbf{M}_y + \mathbf{b})$$

$\mathbf{M}$ und $\mathbf{b}$ sind *weights* bzw. *biases*, φ entspricht der Aktivierungsfunktion. Nachteilig an diesem Verfahren ist, dass durch die beschriebene Transformation der Daten mit jedem Zeitschritt Informationen über die ursprünglichen Eingabedaten verloren gehen. Daher entwickelten Hochreiter und Schmidhuber (1997) die in der vorliegenden Studie verwendete Technik des *long short-term memory* (*LSTM*). Eine *LSTM*-Zelle ist so aufgebaut, dass sie zum einen über die Funktionalität eines 'Kurzzeitgedächtnisses' verfügt, mit welchen Informationen aus vorausgegangenen Zeitschritten gespeichert werden können. Zum anderen kann während dem Training bestimmt werden, welche Informationen davon auf Dauer behalten

('Langzeitgedächtnis') bzw. welche auf Grund ihrer Irrelevanz für die Klassifizierung gelöscht werden sollen. Demnach können mit diesem Verfahren aus den Eingabedaten relevante Informationen extrahiert und diese nach Bedarf beliebig lange gespeichert werden. Dies ermöglicht es auf effektive Art, Muster in Zeitreihen zu erkennen.

2.1.4 Training

Nachdem der Aufbau eines Modells festgelegt wurde, müssen dessen erlernbaren Parameter $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n$, $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$ bzw. $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_n$ durch Training mit den sog. Trainingsdaten bestimmt werden. Diese Trainingsdaten umfassen sämtliche Datensätze, die für das Training ausgewählt wurden, bestehend aus den Eingabedaten $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ und den zugehörigen bekannten Gruppen y_1, y_2, …, y_n.

Zur Beurteilung, wie stark das Modell bei der Vorhersage der Gruppenzugehörigkeit innerhalb der Trainingsdaten von der tatsächlichen Gruppe abweicht, ist eine Verlustfunktion $L(y, \hat{y})$ erforderlich. Dabei repräsentiert y die tatsächliche aus den Trainingsdaten stammende Gruppe und $\hat{y}$ die vom Modell berechnete – und somit vorhergesagte – Gruppe. Es gibt verschiedene Verlustfunktionen L, deren Ausgabe aber jeweils ein Maß für die zuvor beschriebene Abweichung bei einem bestimmten Datensatz ist. Beispielsweise gibt die Verlustfunktion *mean squared error* das durchschnittliche Quadrat der Differenz von y und $\hat{y}$ an. Durch die Quadrierung wird sichergestellt, dass der Unterschied zwischen y und $\hat{y}$ ausschließlich durch einen positiven Zahlenwert dargestellt wird und Abweichungen umso stärker ins Gewicht fallen, je größer sie sind. Um diese Differenz innerhalb der Trainingsdaten mit n Datensätzen als Ganzes zu beurteilen, wird üblicherweise der *average loss* $\langle L \rangle$ berechnet:

$$\langle L \rangle = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{y}_i)$$

Generell ist die Verlustfunktion L umso kleiner, je näher $\hat{y}$ an y liegt. Demnach sollte im Training nach solchen Parametern für $\mathbf{M}$ und $\mathbf{b}$ bzw. $\mathbf{F}$ gesucht werden, bei denen $\langle L \rangle$ für die Trainingsdaten ebenfalls möglichst klein ist. Zur Lösung dieses Problems wird bei *deep learning* häufig ein *gradient descent*-Algorithmus verwendet. Dieser Algorithmus startet mit zufälligen Werten für die Parameter, verändert sie und überprüft anschließend, ob $\langle L \rangle$ größer oder kleiner wird. Das Ziel ist es dabei, $\langle L \rangle$ schrittweise so lange zu verkleinern, bis ein globales Minimum gefunden wird, das durch Veränderung der Parameter nicht mehr weiter minimiert werden kann. Das Maß für die Stärke, mit der der Algorithmus die Parameter verändert, wird als Lernrate (*learning rate*) bezeichnet. Sie sollte während des Trainings nicht konstant sein, sondern je nach den Ergebnissen der Verlustfunktion dynamisch angepasst werden, was u. a. mit dem Optimierungsalgorithmus *ADAM* (Kingma und Ba, 2014) umgesetzt werden kann.

In der Praxis verursacht die Berechnung des *average loss* einen hohen Rechenaufwand, da für jeden Datensatz aus den Trainingsdaten vorab die vorhergesagte Gruppe $\hat{y}$ sowie die davon abhängige Verlustfunktion $L(y, \hat{y})$ bestimmt werden muss. Um die Rechenlast zu minimieren, wird in der Regel auf den *stochastic gradient descent*-Algorithmus zurückgegriffen. Anstatt den gesamten Datensatz für jede Iteration zu verwenden, wird $\langle L \rangle$ nur anhand einer geringen Anzahl an Datensätzen berechnet, die als *batch* bezeichnet wird. Obwohl dabei gewisse Ungenauigkeiten auftreten, bleiben diese in einem akzeptablen Rahmen. Ein Trainingszyklus bzw. eine Trainingsepoche ist abgeschlossen, wenn alle Datensätze aus den Trainingsdaten verarbeitet wurden.

Neben den durch Training erlernbaren Parametern wird die Steuerung des Trainingsalgorithmus zusätzlich durch Hyperparameter bestimmt. Diese werden vor dem Training festge-

legt und umfassen beispielsweise die Anzahl, die Auswahl sowie die Anordnung der genutzten *layers*, die *learning rate* oder die Anzahl an Trainingsepochen. Die Optimierung dieser Hyperparameter ist von großer Bedeutung und bildet somit neben der Vorverarbeitung der Eingabedaten eine weitere Stellschraube zur Erhöhung der Klassifikationsgenauigkeit des Modells.

2.1.5 Validierung

Das fertige Modell soll dazu genutzt werden, neue, bisher unbekannte Eingabedaten korrekt zu klassifizieren. Es ist daher erforderlich, einen kleinen Anteil aller verfügbaren Datensätze nicht für das Training zu verwenden, sondern für eine anschließende Validierung zurückzuhalten. Dadurch kann nach dem Trainings die Klassifizierungsleistung anhand solcher Datensätzen überprüft werden, die nicht aus den Trainingsdaten stammen und dem Modell somit nicht bekannt sind. Diese sog. Validierungsdaten werden als zweite Datenmenge neben den Trainingsdaten benötigt, um einen ersten Eindruck von der Generalisierbarkeit des Modells zu erhalten. Die Rate der korrekt klassifizierten Datensätze aus den Validierungsdaten ist ein möglicher Indikator für diese Generalisierbarkeit und wird daher oft zur Optimierung der Hyperparameter des Modells verwendet.

Als dritte Datenmenge werden Testdaten genutzt, die es ermöglichen, eine abschließende Aussage über die Generalisierbarkeit des Modells zu tätigen. Dies ist deshalb sinnvoll, weil eine Anpassung der Hyperparameter anhand der Klassifizierungsleistung der Validierungsdaten dazu führt, dass Informationen über diese Daten indirekt in das Modell gelangen (*data leakage*). Die Testdaten müssen ähnlich den Validierungsdaten im Vorfeld von den verfügbaren Datensätzen abgespalten werden, dürfen jedoch nur einmalig für eine abschließenden Evaluation des Modells verwendet werden. Ein weiteres Validierungsverfahren, das sich vor allem dann eignet, wenn nur wenige Datensätze verfügbar sind, ist die k-fache Kreuzvalidierung (*k-fold cross-validation*). Hierbei werden alle zur Verfügung stehenden Datensätze zufällig ausgewählt und in k möglichst gleichgroße Teilmengen aufgeteilt. Anschließend erfolgen k Testdurchläufe, bei denen jeweils eine Teilmenge zur Validierung und die restlichen Teilmengen zum Training verwendet werden. Die Wahl dieser Teilmengen wechselt mit jedem Durchlauf, sodass jede davon genau einmal zur Validierung verwendet wird. Dieses Verfahren hat den Vorteil, dass alle zur Verfügung stehenden Datensätze sowohl zum Training als auch zur Validierung des Modells genutzt und somit die Auswirkungen von zufälligen Variationen auf die Zusammensetzung der Trainings- und Validierungsdaten minimiert werden. Nachteilig ist, dass aus diesem Verfahren anstatt eines finalen Modells mehrere unterschiedliche Modelle resultieren und keine Aussage darüber getroffen werden kann, welches der Modelle für eine praktische Anwendung im klinischen Alltag am besten geeignet ist.

Während des Trainings ändert sich die Klassifizierungsleistung des Modells nachdem eine Trainingsepoche durchlaufen wurde. Eine Unteranpassung (*underfitting*) des Modells tritt auf, wenn mit diesem keine hohe Klassifizierungsleistung bei den Trainingsdaten erreicht wird. Dies ist in der Regel dann der Fall, wenn ein Modell nicht komplex genug aufgebaut ist, um ein komplexes System generalisieren zu können. Von Überanpassung (*overfitting*) wird hingegen gesprochen, wenn die Klassifizierungsleistung bei den Trainingsdaten deutlich besser ist als bei den Validierungsdaten. In der Praxis ist *overfitting* von größerer Bedeutung und kann bildlich als 'Auswendiglernen' anstatt 'Generalisierbarkeit' der Trainingsdaten verstanden werden. Die Klassifizierung erfolgt in diesem Fall anhand von *features*, die zufällig in den Trainingsdaten vorhanden, für eine generelle Klassifizierung jedoch irrelevant sind (Vieira et al., 2017). Strategien zur Vermeidung von *overfitting* sind u. a. die Erhöhung der Anzahl an Datensätzen in den Trainingsdaten oder die Reduzierung der Trainingsepochen. Laut den Ergebnissen von Jollans et al. (2019) sind für den Bereich *neuroimaging* mindestens 400 Beobachtungen nötig, um durch klassisches *machine learning* Vorhersagen angemessener Ge-

nauigkeit tätigen zu können. Weitere relevante Einflussfaktoren sind laut den Autoren das Verhältnis von *features* pro Datensatz oder die Effektstärke. Letztgenannte entspricht dem Ausmaß, in dem sich die zu untersuchenden Gruppen in Bezug auf ein bestimmtes *feature* voneinander unterscheiden (Sullivan und Feinn, 2012).

Eine einfache Metrik zur quantitativen Messung der Klassifizierungsleistung ist die Vorhersagegenauigkeit (*accuracy*), die den Anteil an korrekten Klassifizierungen unter allen getroffenen Klassifizierungen angibt. Bei unausgewogenen Datensätzen, in denen eine Gruppe im Vergleich zu den anderen deutlich überrepräsentiert ist, ist die *accuracy* allerdings als Metrik ungeeignet, da ein Modell theoretisch auch eine hohe *accuracy* erreichen kann, indem es ausnahmslos die überrepräsentierte Gruppe vorhersagt. Brodersen et al. (2010) empfohlen daher, die *accuracy* durch die *balanced accuracy* zu ersetzen, die als der Durchschnitt der für jede Gruppe erzielten *accuracy* definiert ist. Eine weitere populäre Metrik bei binären Klassifizierungen ist die sog. Fläche unter der *ROC*-Kurve (*area under the receiver operating characteristic curve, ROC AUC*). Zur Darstellung der *ROC*-Kurve werden in einem Diagramm die Richtig-Positiv-Rate der Vorhersagen als Ordinate und die Falsch-Positiv-Rate als Abszisse eingetragen. Diese Raten bezeichnen die Wahrscheinlichkeit, ein positives Objekt korrekt als positiv bzw. ein negatives Objekt fälschlicherweise als positiv zu klassifizieren. Die *ROC AUC* entspricht der Fläche unter der *ROC*-Kurve, wobei ein Wert von 1 ein perfektes Modell repräsentiert und ein Wert von 0,5 eine rein zufällige Klassifizierung. Eine noch differenziertere Aussage über die Klassifizierungsleistung ermöglichen die in der medizinischen Literatur gebräuchlichen Größen der Sensitivität und der Spezifität. Die Sensitivität gibt an, zu welchem prozentualen Anteil die Krankheit bei tatsächlich Erkrankten erkannt wurde. Im Gegensatz dazu liefert die Spezifität eine Aussage darüber, zu welchem Prozentsatz die Gesunden korrekt als gesund klassifiziert wurden.

Für den einzelnen Probanden ist in der personalisierten Medizin neben der eigentlichen Klassifizierung auch eine Aussage über die Zuverlässigkeit der getroffenen Vorhersage von großer Bedeutung (Jiang et al., 2012). Als mathematisches Äquivalent der Zuverlässigkeit kann die Kalibrierung des Modells herangezogen werden (Nixon et al., 2019). Diese gibt allgemein an, inwiefern die vorhergesagte Wahrscheinlichkeit der Zugehörigkeit zu einer bestimmten Gruppe mit dem tatsächlichen Auftreten übereinstimmt. Zur Bestimmung werden beispielsweise alle Probanden betrachtet, bei denen das Modell eine Wahrscheinlichkeit von 75 bis 80 % für das Vorliegen von IPS angab. Bei einem gut kalibrierten Modell läge dann bei 75 bis 80 % aller Probanden in dieser Subklasse tatsächlich IPS vor. Dies lässt sich grafisch anhand einer Kalibrierungskurve veranschaulichen (vgl. Abbildung 2.2), die zudem dafür genutzt werden kann, die Kalibrierung eines Modells zu optimieren. Die Zuverlässigkeit der berechneten Vorhersagen wurde in der vorliegenden Studie mit dem *Brier score* (Brier, 1950) quantifiziert, der folglich auch der Kalibrierung des Modells diente. Der *Brier score* ist definiert als die mittlere quadratische Abweichung der vorhergesagten Wahrscheinlichkeit aller Gruppen von deren tatsächlichem Auftreten und sollte daher möglichst kleine Werte annehmen.

2.2 Physik der MRT-Bildgebung

Strukturelle MRT-Aufnahmen des Kopfes werden typischerweise erstellt, um die Anatomie der Schädelknochen, der Weichteile inklusive der grauen sowie weißen Substanz und eventuelle Pathologien in Form von Schnittbildern mit einem möglichst hohen Kontrast abzubilden. Eine einzelne dreidimensionale Aufnahme wird als *volume* bezeichnet und besteht aus Voxeln mit Kantenlängen zwischen 0,5 und 4 mm. Funktionelle MRT-Aufnahmen werden verwendet, um die neuronale Aktivität im Gehirn zu untersuchen. Im Gegensatz zur Bildakquisition der

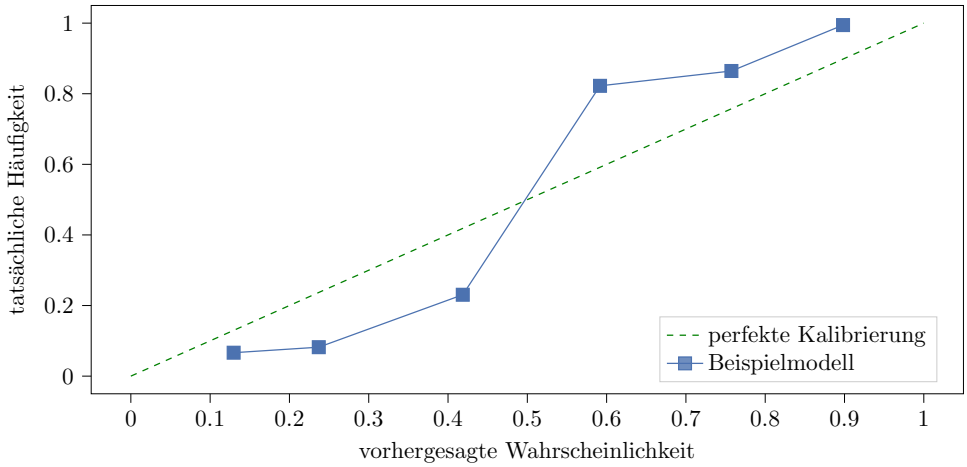

Abb. 2.2: **Kalibrierungskurve eines Beispielmodells.** Das Modell gibt dabei die tatsächliche Häufigkeit für das Vorliegen von IPS in Abhängigkeit von der vorhergesagten Wahrscheinlichkeit an. Probanden mit ähnlichen, vom Modell vorhergesagten, Wahrscheinlichkeiten werden dafür zu einer von insgesamt sechs Subklassen (*bins*) zusammengefasst. Anschließend werden für jede Subklasse die durchschnittliche vorhergesagte Wahrscheinlichkeit und das durchschnittliche tatsächliche Vorhandensein von IPS als Punkt eingetragen. Mit diesen Informationen kann abgelesen werden, ob das Modell die Wahrscheinlichkeit für IPS in bestimmten Subklassen unter- oder überschätzt, sodass zukünftige Vorhersagen entsprechend korrigiert werden können.

strukturellen Aufnahmen werden dabei Serien von Hunderten *volumes* in zeitlich konstanten Abständen von wenigen Sekunden akquiriert. Diffusionsgewichtete MRT-Aufnahmen können u. a. dazu verwendet werden, um Diffusionseinschränkungen oder den Verlauf von Nervenfaserbündel darzustellen. Die Aufnahmen aus allen drei Bildgebungsmodalitäten wurden in der vorliegenden Studie als Eingabedaten der entsprechenden Classifier verwendet. Zum besseren Verständnis dieser Daten handeln die folgenden Abschnitte von den physikalischen Grundlagen der genannten Bildgebungsmodalität sowie deren Besonderheiten.

2.2.1 Strukturelle MRT-Bildgebung

Die MRT-Bildgebung beruht generell hauptsächlich auf den Eigenschaften von Wasserstoffprotonen, die in Form von Wasser fast ubiquitär im menschlichen Körper vorhanden sind. Sie verfügen über einen Drehimpuls (*spin*) und induzieren auf Grund ihrer elektrischen Ladung ein magnetisches Dipolmoment. Vergleichbar mit einem Kreisel sind Wasserstoffprotonen bestrebt, die Lage ihrer Rotationsachsen beizubehalten. Bei Anlage eines starken äußeren Magnetfeldes der Stärke B_0 richten sich diese Rotationsachsen parallel oder antiparallel zum Magnetfeld aus, wobei zum Großteil die energetisch günstigere Parallelposition eingenommen wird und folglich eine gerichtete Längsmagnetisierung M_z entlang der Drehachsen entsteht. Da in dieser Ausrichtung der ursprüngliche Drehimpuls gestört ist, führen die Wasserstoffprotonen eine Präzessionsbewegung durch, die mit dem Torkeln eines Kreisels vergleichbar ist. Diese Bewegung hat eine charakteristische Frequenz ω_0, die als Larmorfrequenz bezeichnet wird. Sie ist zudem von einer für jede Teilchenart spezifischen Konstante γ_0 abhängig und kann durch

$$\omega_0 = \gamma_0 \cdot B_0$$

berechnet werden (Pooley, 2005). In diesem Zustand kann auf die Wasserstoffprotonen Energie übertragen werden, indem sie durch einen Hochfrequenzimpuls mit der Larmorfrequenz ω_0 anregt und somit von der Ausrichtung des Magnetfeldes M_z abgebracht werden.

Nach Abschalten des Hochfrequenzimpulses nehmen die Wasserstoffprotonen wieder ihre vorherige Position ein, was als Relaxation bezeichnet wird. Während der Längsrelaxation geben sie die zuvor aufgenommene Energie in Form eines magnetischen Impulses wieder ab und induzieren in den Messspulen des Scanners eine Spannung, auf der die Bildgebung beruht. Bei der Querrelaxation findet eine Desynchronisierung der Präzessionsbewegung ohne Energieabgabe statt.

Mathematisch werden den Relaxationsvorgängen gewebespezifische Zeitkonstanten zugeordnet, die bei der Längsrelaxation als Spin-Gitter-Relaxationszeit oder T1 und bei der Querrelaxation als Spin-Spin-Relaxationszeit oder T2 bezeichnet werden. Je nachdem, ob eine Aufnahmesequenz T1- oder T2-gewichtet ist, werden Gewebe auf Grund dieser unterschiedlichen Relaxationszeiten in der rekonstruierten Aufnahme unterschiedlich hyper- bzw. hypointens dargestellt, sodass verschiedene Fragestellungen spezifisch analysiert werden können. Durch Überlagerung des Magnetfeldes mit zusätzlichen, in eine Richtung stärker werdenden Magnetfeldern (Gradientenfeldern) kann der Hochfrequenzimpuls nur noch Protonen innerhalb eines schmalen Bereichs anregen. Dieser örtlich-magnetische Gradient bewirkt also eine Abnahme der Larmorfrequenz γ_0 entlang des Gradienten. Somit kann mit γ_0 auch eine Ortscodierung erfolgen, indem Gradientenfelder entlang aller drei Raumachsen angelegt werden. Zur Minimierung des Bildrauschens werden unterschiedliche Pulssequenzen verwendet, in denen die Wasserstoffprotonen durch Hochfrequenzimpulse angeregt werden, das magnetische Feld verändert und das gemessene Signal gemittelt wird. Die Zeit zwischen zwei Anregungen wird als Repetitionszeit (*repetition time, TR*) und die Zeit zwischen Anregung und Signalaufnahme als Echozeit (*echo time, TE*) bezeichnet. Bei T1-gewichteten Aufnahmen sind *TR* und *TE* im Vergleich zur Relaxationszeit kurz und bei T2-gewichteten Aufnahmen vergleichsweise lang.

2.2.2 Funktionelle MRT-Bildgebung

Die funktionelle MRT-Bildgebung ist ein Verfahren zur Darstellung von Durchblutungsveränderungen des Gehirns, die als Surrogatmarker für neuronale Aktivität verstanden werden können (Logothetis und Wandell, 2004). Die physikalische Grundlage dafür bilden die unterschiedlichen magnetischen Eigenschaften von Hämoglobin, abhängig davon, ob es oxygeniert oder desoxygeniert ist (Thulborn et al., 1982). Somit ergibt sich während der Bildakquisition eine andere Signalintensität für sauerstoffarmes Blut im Vergleich zu sauerstoffreichem Blut. Dieser Effekt wird als *blood oxygen level dependent (BOLD) effect* bezeichnet und das gemessene Signal entsprechend als *BOLD*-Signal. Der zeitliche Verlauf des reinen *BOLD*-Signals nach Einsetzen von neuronaler Aktivität wird durch die hämodynamische Antwortfunktion (*hemodynamic response function*) beschrieben (vgl. Abbildung 2.3). Sie zeigt, dass das Signal ca. 5 s nach Einsetzen neuronaler Aktivität seinen Höhepunkt erreicht und nach ca. 20 s wieder auf den Ausgangswert zurückfällt (Huettel et al., 2001). Die Reaktion tritt also zeitlich deutlich verzögert auf, da als physiologische Antwort auf den erhöhten Sauerstoffbedarf der Neuronen nach Aktivierung zuerst eine Dilatation der versorgenden Arteriolen mit Erhöhung des zerebralen Blutflusses erfolgt – der für das *BOLD*-Signal entscheidende Abfall von desoxygeniertem Hämoglobin geschieht erst danach. Auf Grund dieser Verzögerung ist es folglich zur Untersuchung von neuronalen Aktivitäten ausreichend, wenn die Akquisitionszeit der einzelnen *volumes* einige Sekunden beträgt. Im Vergleich zur Elektroenzephalographie bietet die funktionelle MRT also eine hohe räumliche und eine niedrige zeitliche Auflösung.

Die physikalischen Grundlagen der Bildakquisition sind mit denen der strukturellen MRT vergleichbar, jedoch wird eine Pulssequenz benötigt, mit der einzelne *volumes* innerhalb von Sekunden erfasst werden können. Dies ist mit der sog. *echo planar imaging (EPI)*-Sequenz möglich und entsprechend werden mit dieser Sequenz erzeugte *volumes* als *EPI-volumes* be-

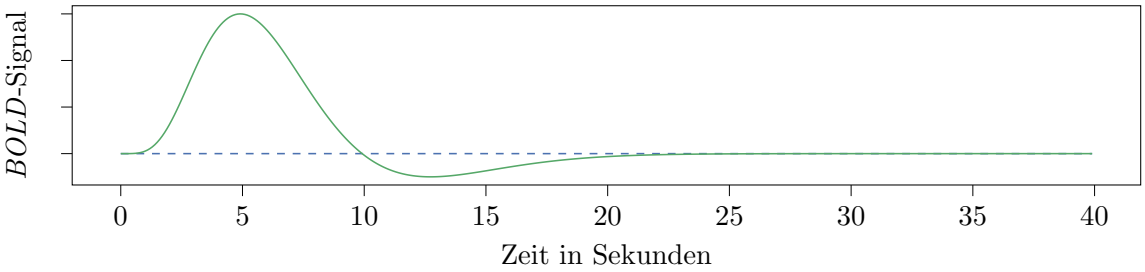

Abb. 2.3: ***Hemodynamic response function* des *BOLD*-Signals nach Stimulus.**

zeichnet. Es erfolgt dabei typischerweise pro *EPI-volume* nur eine einzige Anregung durch die Hochfrequenzimpulse, sodass die dafür benötigte Bildakquisitionszeit der Repetitionszeit der Pulssequenz entspricht. Die einzelnen Zeitpunkte, zu denen während der Untersuchung die Serien an *EPI-volumes* akquiriert werden, werden als *time points* bezeichnet und deren Gesamtanzahl entspricht der Anzahl aller akquiriert *EPI-volumes*. Als weitere Besonderheit wird für die Sequenz eine sog. T2*-Gewichtung verwendet. Diese basiert zwar auf der T2-Gewichtung, ist jedoch besonders sensibel für solche Inhomogenitäten des Magnetfeldes, die durch eine Zunahme an desoxygeniertem Hämoglobin verursacht werden. Dies ermöglicht eine präzisere Detektion des relevanten *BOLD*-Signals. Im Vergleich zur strukturellen MRT haben *EPI-volumes* eine geringere Auflösung, sind anfälliger für Bildartefakte und zeigen einen geringeren Kontrast zwischen den Gewebetypen.

2.2.3 Diffusionsgewichtete MRT-Bildgebung

Die klassische diffusionsgewichtete MRT-Bildgebung wird üblicherweise dazu verwendet, die ungerichtete Bewegung von Wassermolekülen im Gehirn darzustellen, um somit z. B. Rückschlüsse auf unterversorgtes Gewebe ziehen zu können. Im Gegensatz dazu ist bei der Diffusions-Tensor-Bildgebung der Verlauf von Nervenfaserbündeln bzw. die anatomische Konnektivität von Interesse. Da die Wassermoleküle nicht frei durch die Zellmembran der Neuronen diffundieren können, verläuft ihre Bewegung intra- sowie extrazellulär zum Teil gerichtet entlang der Axone. Diese axiale Bewegungskomponente ist im Vergleich zum Anteil radialer und perpendikulärer Bewegungen umso größer, je dichter die Axone beieinanderliegen (Le Bihan, 2003), sodass anhand dieser Diffusionsbewegung der Verlauf von größeren Nervenleitungsbahnen berechnet werden kann.

Bei der *DTI* gibt es unterschiedliche Metriken, um Menge und Richtung der Diffusion darzustellen. In der vorliegenden Studie wurden dabei die fraktionale Anisotropie, die radiale Diffusionsfähigkeit, die mittlere sowie die axiale Diffusivität berechnet. Die fraktionale Anisotropie ist gut dazu geeignet, die weiße Substanz von anderen Gewebetypen zu unterscheiden und Auffälligkeiten innerhalb dieses Gewebes hervorzuheben. Hingegen lässt sich mit der mittleren Diffusivität der Liquor cerebrospinalis gut darstellen. Mit der radialen Diffusionsfähigkeit und der axialen Diffusivität können die Kontraste zwischen allen Gewebetypen abgebildet werden (vgl. Abbildung 2.4).

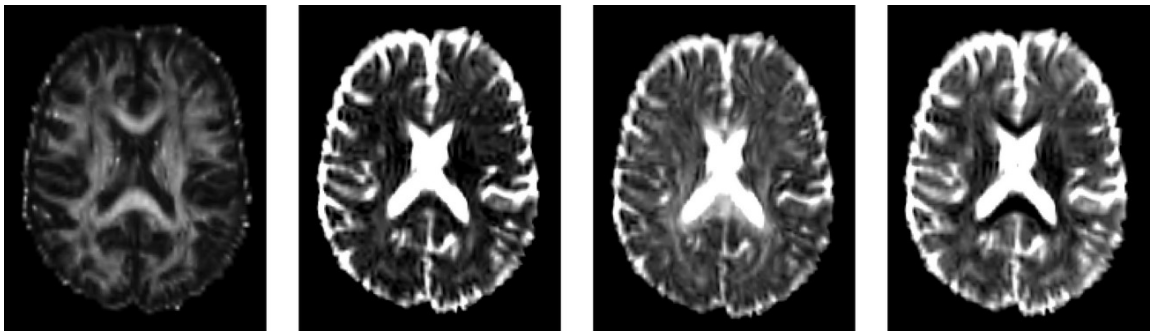

Abb. 2.4: **Unterschiedliche *DTI*-Metriken.** Fraktionale Anisotropie, mittlere sowie axiale Diffusivität und radiale Diffusionsfähigkeit (von links nach rechts).

2.3 Methoden der Bildverarbeitung

Die Variabilitäten von MRT-Aufnahmen, die nicht durch Probanden selbst bedingt sind, werden von Fortin et al. (2018) als *scanner effects* bezeichnet und erschweren den Vergleich der Aufnahmen untereinander. *Scanner effects* entstehen beispielsweise durch Verwendung unterschiedlicher Modelle von MRT-Scannern sowie durch uneinheitliche Erfassungsprotokolle (*imaging protocols*), durch Signalrauschen bzw. Artefakte oder Bewegungen von Probanden während der Bildakquisition, weshalb zwangsläufig alle Aufnahmen davon betroffen sind. Für eine bestmögliche Vergleichbarkeit zwischen den Probanden ist daher eine umfangreiche Vorverarbeitung der Aufnahmen notwendig.

2.3.1 NIfTI-Standard

Zur Archivierung und zum Austausch medizinischer Bilddaten stehen unterschiedliche Dateiformate zur Verfügung. Der sog. DICOM-Standard ist herstellerübergreifend in einer großen Anzahl bildgebender Medizinprodukte implementiert und ermöglicht es, neben den reinen digitalen Bilddaten auch korrespondierende Metainformationen, wie beispielsweise Patienten- oder Geräteinformationen, abzuspeichern. Da dieses Format jedoch nicht für die weitere Bildverarbeitung entwickelt wurde, entstand zu diesem Zweck der sog. NIfTI-Standard (Whitcher et al., 2011). Dieser bietet ebenfalls die zuvor genannten Vorzüge und ermöglicht auf Grund seiner Kompatibilität mit vielen Softwarepaketen einen einfachen Umgang mit Bilddaten aus dem Bereich der neuronalen Bildgebung (*neuroimaging*). Im DICOM-Format bereitgestellte MRT-Aufnahmen können mit Softwaretools wie beispielsweise dcm2niix (X. Li et al., 2016) zu NIfTI-Dateien konvertiert werden.

Die rekonstruierten Bildinformationen der MRT-Messungen liegen in Abhängigkeit von der gewählten Bildgebungsmodalität als drei- oder vierdimensionales Array vor. Strukturelle Aufnahmen umfassen lediglich drei Dimensionen, die die Breite, die Höhe und die Tiefe dieser Aufnahmen in Voxel repräsentieren. Für funktionellen sowie diffusionsgewichteten Aufnahmen besteht zusätzlich zu den drei räumlichen Dimensionen eine zeitliche Dimension, da bei diesen Bildgebungsmodalitäten MRT-Messungen zu mehreren Zeitpunkten (*time points*) durchgeführt werden. Die einzelnen Elemente eines Arrays entsprechen den Signalintensitäts- bzw. 'Helligkeitswerten' der Voxel, die wiederum die Protonendichte und unterschiedliche Relaxationszeiten verschiedener Gewebearten widerspiegeln (vgl. Abschnitt 2.2.1). Zusätzlich wird für die exakte Bildrekonstruktion eine 4×4-Transformationsmatrix (*affine matrix*) benötigt (vgl. Abschnitt 2.3.3), die in den Metainformationen gespeichert ist. In ihr sind Angaben zur räumlichen Orientierung in einem Referenzraum enthalten, sodass eine Zuordnung von anterior, posterior, superior, inferior sowie links und rechts auf den Probanden, unabhängig von seiner Liegeposition im MRT-Scanner, möglich ist.

2.3.2 Brain Extraction

Die Gehirnextraktion (*brain extraction*) bezeichnet die Entfernung nichtzerebralen Gewebes aus Schnittbildern. Dies verbessert die Ergebnisse der nachfolgenden Bildverarbeitungsschritte (Leung et al., 2011) und verringert die Komplexität sowie Größe der Datensätze, wodurch weitere Analysen weniger rechenintensiv werden. Der Goldstandard für die Durchführung dieses Verfahrens ist eine manuelle Extraktion des Gehirns durch einen Experten (Boesen et al., 2004), was jedoch auf Grund des damit verbundenen Zeitaufwandes keine praktikable Lösung bei der Untersuchung einer großen Anzahl an Aufnahmen darstellt. Im Folgenden werden daher zwei automatisierte Methoden vorgestellt, die in der vorliegenden Studie verwendet wurden.

CAT12 (Gaser und Dahnke, 2016) verwendet zur Identifizierung des Gehirns ein dafür angepasstes *region growing*-Verfahren. Bei dieser Methode wird ein Bereich geschätzt, der sich im Gehirn befindet. Anschließend werden benachbarte Regionen mit diesem Bereich mittels spezifischer Kriterien verglichen, sodass beispielsweise anhand von Voxel-Intensitäten festgestellt werden kann, ob diese Regionen ebenfalls zum Gehirn gehören. Sollte dies der Fall sein, so verschmelzen sie mit dem ursprünglichen Bereich (Park und Lee, 2009). Die somit gewonnenen Erkenntnisse werden zur Erstellung von *tissue probability maps* verwendet, die jedem Voxel der ursprünglichen Aufnahme eine Wahrscheinlichkeit für beispielsweise graue Substanz, weiße Substanz oder Liquor cerebrospinalis zuordnen. In Kombination mit weiteren *tissue probability maps* nach einem Verfahren von Ashburner und Friston (2005) kann anschließend eine binäre *brain mask* abgeschätzt werden. Diese ist ähnlich einer einzelnen *tissue probability map* aufgebaut, jedoch wurden die zuvor darin enthaltenen Wahrscheinlichkeiten für das Vorliegen eines bestimmten Hirngewebes anhand eines Schwellenwertes durch 0 oder 1 ersetzt. Das extrahierte Gehirn wird schließlich durch den Abgleich der *brain mask* mit der ursprünglichen Aufnahme berechnet.

Bei Verwendung des Softwarepakets FSL BET (Jenkinson et al., 2005; S. M. Smith, 2002) wird zuerst ein Histogramm aller Voxel-Intensitäten erstellt. Unter Ausschluss der kleinsten und größten 2 % aller Werte (Ausreißer) werden Maximal- sowie Minimalwert bestimmt. Anschließend wird ein Schwellenwert l_{thresh} zur Abgrenzung zerebralen Gewebes vom Hintergrund der Aufnahme abgeschätzt. Anhand aller Voxel-Intensitäten größer als l_{thresh} können somit grob der Mittelpunkt (*center of gravity*) sowie der Radius des Gehirns bestimmt werden. Eine aus verbundenen Dreiecken bestehende tesselierte Kugel mit diesem Mittelpunkt und halbem Radius wird danach so lange nach außen hin verformt, bis sie die Oberfläche des Gehirns abbildet. Dabei werden die Dreiecke schrittweise in kleinere Dreiecke aufgeteilt und deren Vertexpositionen angepasst, sodass sie im Verbund eine Form bilden, deren Inhalt dem extrahierten Gehirn entspricht. Ähnlich der bei CAT12 verwendeten Methode entsteht somit eine *brain mask*, aus der hervorgeht, in welchen Bereichen der Aufnahme sich zerebrales Gewebe befindet. In Kombination mit der Originalaufnahme kann so schließlich das extrahierte Gehirn rekonstruiert werden (vgl. Abbildung 2.5).

2.3.3 Bildregistrierung

Das primäre Ziel der Bildregistrierung bzw. Normalisierung ist die Angleichung aller Aufnahmen einer oder mehrerer Bildgebungsmodalitäten aneinander, sodass Voxel mit identischen Koordinaten bei allen Probanden den gleichen anatomischen Ort darstellen und folglich die Vergleichbarkeit zwischen den Probanden erleichtert wird. Des Weiteren wird die sog. Koregistrierung verwendet, um Aufnahmen unterschiedlicher Bildgebungsmodalitäten aneinander anzugleichen, sodass ein bestimmter anatomischer Ort in einer beliebigen Bildgebungsmodalität auch in Aufnahmen anderer Bildgebungsmodalitäten exakt aufgefunden

Abb. 2.5: ***Brain extraction* mit BET.** Die Originalaufnahme (links) wird mit der *brain mask* (Mitte) multipliziert, um das extrahierte Gehirn (rechts) zu erhalten.

werden kann. Diese Angleichungen geschehen durch lineare sowie nichtlineare Bildtransformationen der Aufnahmen. Lineare Transformationen (*affine transformations*) für Translation, Rotation, Skalierung und Scherung umfassen jeweils drei Freiheitsgrade (*degrees of freedom*) entlang drei Raumachsen (vgl. Abbildung 2.6), wohingegen nichtlineare Transformationen (*warps*) theoretisch beliebig viele Freiheitsgrade beinhalten, sodass auch Aufnahmen mit lokalen Unterschieden exakt aufeinander ausgerichtet werden können. Die Durchführung der linearen oder nichtlinearen Transformation erfolgt durch Festlegung einer Referenzaufnahme, an die die restlichen Aufnahmen durch Bildtransformationen möglichst exakt angeglichen werden. Als solche Referenzaufnahme wurde in der vorliegenden Studie die neuroanatomische Vorlage MNI152 des Montreal Neurological Institute verwendet, bei der 152 strukturelle T1-gewichtete MRT-Aufnahmen des Kopfes von gesunden Probanden zu einer einzelnen, möglichst repräsentativen Aufnahme zusammengefasst wurden (Evans et al., 1993; Fein et al., 2004). Mittels Bildregistrierung anhand dieser Referenzaufnahme werden entsprechende Aufnahmen in den sog. MNI152-Standardraum überführt, der u. a. durch eine feste Auflösung charakterisiert ist, was für die spätere Anwendung von dualer Regression und verschiedenen *deep learning*-Verfahren vorteilhaft ist.

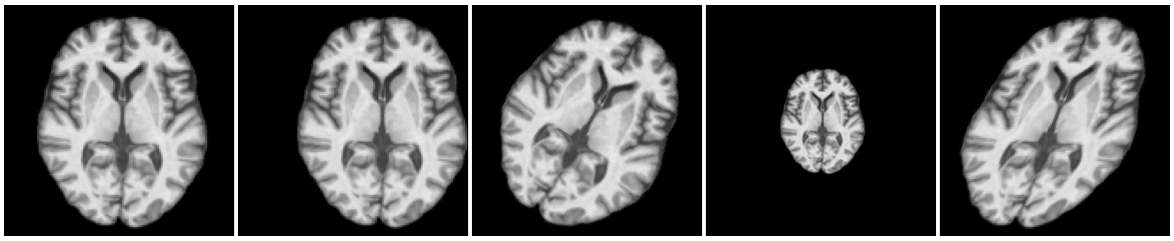

Abb. 2.6: **Affine Transformationen einer zweidimensionalen Aufnahme.** Originalaufnahme, Translation, Rotation, Skalierung und Scherung (von links nach rechts).

Die Umsetzung der linearen Bildregistrierung wird im Folgenden beispielhaft anhand der Funktionsweise des Softwarepakets FSL FLIRT (Jenkinson und Smith, 2001; Jenkinson et al., 2002) erklärt. Bei deren Anwendung wird die Qualität der Bildregistrierung unter Verwendung einer Verlustfunktion C bewertet und anschließend nach der Transformation T^* gesucht, für die C ein globales Minimum erreicht, das der bestmöglichen Ausrichtung einer Aufnahme entspricht. Bei den Berechnungen dieser Transformationen nimmt der Rechenaufwand mit steigender Anzahl an Freiheitsgraden überproportional zu, sodass T^* nur schrittweise bestimmt werden kann. Dazu wird der Bereich des vermuteten Minimums von C abgeschätzt, indem Transformationen mit nur wenigen Freiheitsgraden bei zusätzlich verringerter Auflösung der Aufnahme durchgeführt und bewertet werden. Erst danach kann die Anzahl an Freiheitsgraden und die Auflösung wieder erhöht werden, um genauere Parame-

ter für T^* berechnen zu können. Da die Aufnahmen aus scharf voneinander abgetrennten Voxeln bestehen und somit diskretisiert sind, werden außerdem Interpolationsverfahren wie beispielsweise trilineare Interpolation benötigt, um unbekannte Helligkeitswerte in der transformierten Aufnahme zu berechnen.

Eine globale affine Transformation T^* kann mathematisch mit einer 4×4-Transformationsmatrix beschrieben werden (Wenckebach, 2005) und dient im Rahmen der Bildregistrierung der Angleichung einer Aufnahme an die Referenzaufnahme. Die Voxelkoordinaten x_1, y_1 und z_1 der Aufnahme entsprechen nach Multiplikation mit T^* den Voxelkoordinaten x, y und z der Referenzaufnahme:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = T^* \cdot \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} \quad \text{mit } T^* = \begin{bmatrix} a_{00} & a_{01} & a_{02} & t_0 \\ a_{10} & a_{11} & a_{12} & t_1 \\ a_{20} & a_{21} & a_{22} & t_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Dabei beschreiben die Parameter a_{00} bis a_{22} die Rotations-, die Skalierungs- sowie die Scherungskomponenten und t_0 bis t_2 die Translationskomponente.

Mittels nichtlinearer Bildregistrierung können schließlich auch lokale Transformationen innerhalb der Aufnahmen erfolgen, die häufig den Vergleich zwischen Probanden vereinfachen (Klein et al., 2009). Vor Anwendung nichtlinearer Transformationen sollte jedoch jeweils eine lineare Bildregistrierung durchgeführt werden, damit die geplanten Transformationen durch sinnvolle Ausgangswerte schneller und präziser ermittelt werden können (Warfield et al., 1999). Zudem sind auf Grund der sehr hohen Anzahl an Freiheitsgraden bei nichtlinearen Transformationen zusätzliche Maßnahmen zur Regularisierung (*regularization*) notwendig, um die Topologie des Gehirns möglichst gut zu erhalten. Die Regularisierung ist in diesem Kontext ein Maß dafür, wie stark die Geometrie der Aufnahme verändert werden darf, wobei mit steigender *regularization* die Transformationen vermehrt über die Aufnahme verteilt und weniger lokal durchgeführt werden.

Die Parameter einer nichtlinearen Transformation sind in der Regel so umfangreich, dass sie nicht durch Transformationsmatrizen, sondern mit Hilfe sog. *warp fields* oder *deformation fields* beschrieben werden. Diese enthalten beispielsweise für jeden Voxel der Aufnahme einen Vektor, der Richtung und Stärke der Transformation für einen mit der Aufnahme korrespondierenden lokalen Bereich repräsentiert. Um Informationsverluste durch mehrfache Interpolationsvorgänge zu vermeiden, wird anhand der Transformationsmatrix der linearen Transformation und der *warp fields* bzw. *deformation fields* der nichtlinearen Transformation eine daraus resultierende Transformation berechnet, sodass nach Anwendung dieser auf die Aufnahme nur eine einzige Interpolation notwendig ist.

2.3.4 Voxel-basierte Morphometrie

Bei der Voxel-basierten Morphometrie handelt es sich um ein Verfahren, mit dem nach Bildregistrierung in einen Standardraum der Anteil grauer Substanz, weißer Substanz und Liquor cerebrospinalis in jedem Voxel einer Aufnahme abgeschätzt wird (Ashburner und Friston, 2000). Diese Methode ist folglich gut zur Feststellung subtiler hirnmorphologischer Abweichungen bei strukturellen MRT-Aufnahmen zwischen verschiedenen Patientenkollektiven oder einzelnen Probanden geeignet (Scarpazza und De Simone, 2016). Die Voxel-basierte Morphometrie stellt eine Weiterentwicklung der konventionellen Hirnmorphometrie dar, in der lediglich eine *region of interest* von den restlichen Hirnanteilen abgegrenzt und vermessen wird.

Das mit CAT12 durchgeführte Verfahren der Voxel-basierten Morphometrie verwendet zur voxelweisen Abschätzung lokaler Gewebekonzentrationen der grauer Substanz eine in SPM12 (Ashburner et al., 2014) integrierte Methode von Ashburner und Friston (2005), die zwei unterschiedliche Ansätze zur Gewebssegmentierung vereint. In einem ersten Ansatz werden die Aufnahmen mittels Bildregistrierung in einen Standardraum überführt, für den a priori *tissue probability maps* vorhanden sind. Diese enthalten unabhängig von der jeweiligen Aufnahme für jeden Voxel im Standardraum die Wahrscheinlichkeiten, dass sich an einer bestimmten Stelle ein bestimmter Hirngewebetyp befindet. Mittels nichtlinearer Transformationen wird die Aufnahme den *tissue probability maps* angepasst, sodass Informationen zur Gewebeverteilung anhand der durchgeführten Transformationen abgeleitet werden können. In einem zweiten Ansatz erfolgt die Segmentierung der Aufnahme direkt durch ein Schwellenwertverfahren ausgehend von den Voxel-Intensitäten, sodass Voxel entsprechend ihrer Gewebetypen zu zusammenhängenden Regionen zusammengefasst werden. Mit diesem Verfahren ist es zudem möglich, Bildartefakte durch Inhomogenitäten im Magnetfeld während der Bildakquisition zu erkennen und zu korrigieren.

Die somit generierte Ausgabe wird anschließend für weitere Verarbeitungsschritte genutzt. Anders als unter Verwendung von SPM12 wird in CAT12 die finale Segmentierung nach einer Methode von Rajapakse et al. (1997) berechnet, die unabhängig von *tissue probability maps* funktioniert und für die Korrektur von Partialvolumeneffekten optimiert wurde (Tavares et al., 2020), sodass die Gewebeverteilung einzelner Voxel, die mehrere Gewebetypen enthalten, besser abgeschätzt werden kann. Schließlich wird die segmentierte Aufnahme mittels nichtlinearer Transformation in einen Standardraum (z. B. MNI152) überführt, wobei die Aufnahme dabei so moduliert wird, dass die jeweiligen Anteile an grauer Substanz, weißer Substanz und Liquor cerebrospinalis ähnlich wie in der ursprünglichen Aufnahme verteilt sind. Im Hinblick auf einen Vergleich der Aufnahmen miteinander bietet die vorausgegangene Segmentierung von grauer Substanz, Normalisierung und *modulation* den Vorteil, dass *scanner effects*, wie beispielsweise Bildrauschen, weitgehend eliminiert werden und somit die Klassifizierung der Aufnahmen überwiegend anhand hirnstruktureller Eigenschaften der Probanden erfolgen kann.

2.3.5 Bewegungskorrektur

Bereits kleine Bewegungen des Kopfes können während der Bildakquisition bei funktionellen oder diffusionsgewichteten MRT-Aufnahmen dazu führen, dass sich gleiche Strukturen eines Probanden an unterschiedlichen Voxelkoordinaten innerhalb seiner *EPI-volumes* befinden. Diese Bewegungsartefakte können beispielsweise in einer funktionellen Aufnahme in der späteren Auswertung zu falschen Korrelationen einzelner Voxel und somit zu einer ungenauen Bestimmung der funktionellen Konnektivität führen (Power et al., 2012). Power et al. (2015) zeigten dazu in einer Studie, dass bei Probanden, die sich vermehrt bewegten, übermäßig starke Korrelationen für nah beieinanderliegende Hirnregionen und vergleichsweise schwache Korrelationen für weit auseinanderliegende Hirnregionen vorgetäuscht wurden. Zur Vermeidung dieser Problematik sollten Probanden während der gesamten Akquisitionszeit möglichst still liegen und etwaige Bewegungsartefakte softwaregestützt korrigiert werden. Dieser Schritt ist insbesondere bei Patienten mit IPS von großer Bedeutung, da der Ruhetremor ein typisches Symptom darstellt, das zu einer vermehrten Bewegung betroffener Probanden führt und folglich als möglicher Confounder (vgl. Unterkapitel 2.5) bei der späteren Klassifizierung berücksichtigt werden muss.

Die automatisierte Bewegungskorrektur kann u. a. mit dem Softwarepaket FSL MCFLIRT (Jenkinson et al., 2002) umgesetzt werden. Dabei wird ein *EPI-volume* aus der Mitte der Zeitreihe als Referenzaufnahme festgelegt. Vergleichbar mit FSL FLIRT (vgl. Abschnitt 2.3.3)

wird diese Aufnahme zur Bildregistrierung der anderen *EPI-volumes* eines Probanden verwendet, wobei angenommen wird, dass die Korrektur nur eine Transformation von wenigen Millimetern umfasst. Außerdem können die Freiheitsgrade dieser Transformation auf sechs begrenzt werden, da Kopfbewegungen allein durch Translation und Rotation der Aufnahme korrigierbar sind. Zur Abschätzung der dafür verwendeten Korrekturparameter wird zuerst eine Transformationsmatrix mit nur geringer Genauigkeit abgeschätzt. Die dabei erhobenen Parameter dienen als Ausgangswerte für zwei weitere Abschätzungen mit engeren Toleranzen. Nach Durchführung der somit bestimmten Transformation wird dieses Verfahren bei allen übrigen *EPI-volumes* des Probanden wiederholt.

2.3.6 Intensity Normalization

Im Gegensatz zur Computertomographie existiert bei MRT-Aufnahmen auf Grund von Inhomogenitäten des magnetischen Feldes oder wechselnden Aufnahmemodalitäten keine standardisierte Skala für die Intensitätswerte, sodass gleiche Gewebetypen auf unterschiedlichen Aufnahmen nicht durch gleiche Voxel-Intensitäten abgebildet werden (Nyúl und Udupa, 1999). Die Transformation aller Intensitätswerte auf eine standardisierte Skala wird im Kontext von *machine learning* auch als *feature scaling* bezeichnet und verbessert die Vorhersagegenauigkeit fast aller Modelle (Géron, 2019). Eine dafür gebräuchliche Methode ist *rescaling*, die alle Werte x aus den Eingabedaten in das Intervall $[0; 1]$ überführt. Der normalisierte Wert x' wird dabei durch

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

berechnet, wobei $\min(x)$ und $\max(x)$ kleinster bzw. größter Wert aller x sind.

Eine populäre Methode zur Normalisierung von *EPI-volumes* zwischen den Probanden ist *grand mean scaling*. Dabei werden die Voxel-Intensitäten x aller *EPI-volumes* über alle *time points* mit einem probandenspezifischen Faktor f_i multipliziert, sodass die mittlere Intensität einem konstanten Wert k entspricht, der somit bei allen Probanden identisch ist.

$$x' = f_i x \quad \text{mit } f_i = \frac{k}{\text{mean}(x)}$$

Dies bewirkt, dass auch das arithmetische Mittel der Aufnahmen bei allen Probanden gleich ist und sich die Varianz zwischen den Probanden reduziert, wodurch sich in der späteren statistischen Analyse der Datensätze die Trennschärfe verbessert (Jenkinson und Chappell, 2018).

2.3.7 Spatial Smoothing

Spatial smoothing kann angewendet werden, um das Signal-Rausch-Verhältnis in funktionellen MRT-Aufnahmen zu verbessern und Artefakte, die bei der Bildregistrierung entstanden sind, zu reduzieren (Maas und Renshaw, 1999). Dies ist notwendig, da die Intensitätsschwankungen des *BOLD*-Signals nur diskret sind und folglich kaum vom Bildrauschen unterschieden werden können. Implementiert wird dieses Verfahren meistens durch Weichzeichnung der Aufnahme mit einem Gauß-Filter (Lindquist et al., 2010), dessen Stärke durch die Halbwertsbreite (*full width at half maximum*, *FWHM*) festgelegt wird. Der neue Intensitätswert eines jeden Voxels wird dabei in Abhängigkeit von den Intensitätswerten seiner benachbarten Voxel berechnet. In Anlehnung an eine Gauß-Verteilung werden Intensitätswerte nahe gelegener Voxel stärker gewichtet als solche, die weiter entfernt liegen. Mit steigender Halbwertsbreite nimmt die Gewichtung entfernter Voxel, und folglich die Stärke der Weichzeichnung, zu (vgl. Abbildung 2.7). Feines Bildrauschen der Aufnahme wird somit zu Ungunsten der Auflösung reduziert.

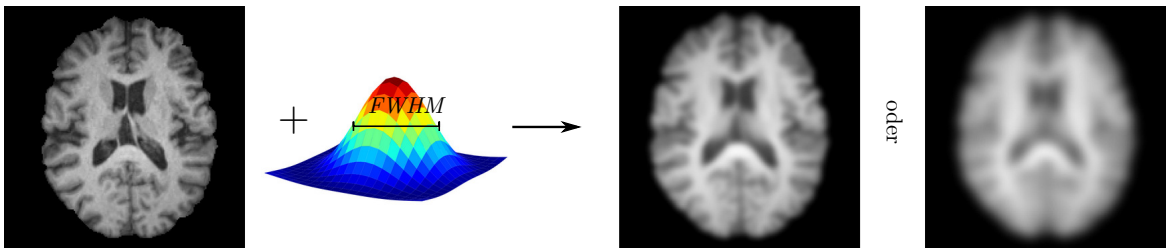

Abb. 2.7: ***Spatial smoothing* mit einem Gauß-Filter.** Als *FWHM* wurden 5 (links) oder 10 mm (rechts) verwendet.

Die ideale Größe der Halbwertsbreite kann im Vorfeld meistens nicht festgelegt werden (Hopfinger et al., 2000), jedoch empfehlen einige Autoren einen Bereich von ungefähr 4 bis 8 mm bei einer Voxelgröße von $2 \times 2 \times 2\,\mathrm{mm}^3$ (Mikl et al., 2008; Chen und Calhoun, 2018).

2.3.8 Temporal Filtering

Da die gesuchten *BOLD*-Signale bei funktionellen MRT-Aufnahmen zu einem großen Teil von zyklisch schwankenden Störsignalen überlagert sind, kann *temporal filtering* dazu genutzt werden, diese ungewollten Effekte herauszufiltern und somit das Signal-Rausch-Verhältnis zu verbessern (Ngan et al., 2000). Die Grundlage dafür ist der relativ schmale Frequenzbereich der *BOLD*-Signale, wohingegen Signale nichtneuronaler Herkunft in einem anderen Bereich anzutreffen sind. Physiologische Signalschwankungen durch Atmung und Herzaktionen treten mit einer Frequenz von über 0,1 Hz auf (Cordes et al., 2001), eine Form des durch den Scanner verursachten Rauschens – der sog. *scanner drift* – hat eine Frequenz von unter 0,01 Hz (A. M. Smith et al., 1999), während das zur Darstellung der funktionellen Konnektivität wichtige *BOLD*-Signal eine Frequenzbreite zwischen 0,01 und 0,1 Hz umfasst (Biswal et al., 1995; Cordes et al., 2001). Es ist dabei gängige Praxis, Signale mit niedriger Frequenz durch Anwendung eines Hochpass-Filters zu unterdrücken, da diese für die weiteren Analysen besonders störend sind (A. T. Smith et al., 2007).

FSL FEAT (Woolrich et al., 2001) bietet dafür einen speziellen Hochpass-Filter, der auf einer Methode von Marchini und Ripley (2000) basiert. Dabei werden Signalverläufe bzw. Zeitreihen der Voxel durch Extraktion der mit *response* bezeichneten Schwankungen der Voxel-Intensitäten im zeitlichen Verlauf gebildet. Des Weiteren werden dazu korrespondierende nichtlineare Anpassungslinien erstellt, indem die zuvor berechneten Signalverläufe durch lokale Mittelung mit einem Gauß-Filter geglättet werden. Diese Anpassungslinie repräsentiert damit die unerwünschten niedrigen Frequenzen, sodass sie von ihren entsprechenden Signalverläufen subtrahiert und die *EPI-volumes* schließlich neu rekonstruiert werden (vgl. Abbildung 2.8).

2.4 Funktionelle Konnektivität

Der Begriff funktionelle Konnektivität bezeichnet die zeitliche Korrelation neuronaler Aktivitäten in unterschiedlichen Hirnregionen (Horwitz, 2003). Sie kann erfasst werden, indem die Zeitreihe eines Voxels mit den Zeitreihen aller anderen Voxel korreliert wird – mit zunehmender Stärke der Korrelation steigt auch die Wahrscheinlichkeit einer Konnektivität zwischen den durch die Voxel repräsentierten Hirnregionen. In mehreren Studien wurde gezeigt, dass bei Patienten mit IPS die funktionelle Konnektivität des Gehirns auf Gruppenniveau verändert ist (Helmich et al., 2010; Sharman et al., 2013; Baudrexel et al., 2011). Hier ergibt sich ein interessanter Ansatzpunkt für die Entwicklung eines Biomarkers. Zur Erfassung dieser Konnektivität sind vor allem Aufnahmen im Ruhezustand (*resting-state*) geeignet (Eickhoff

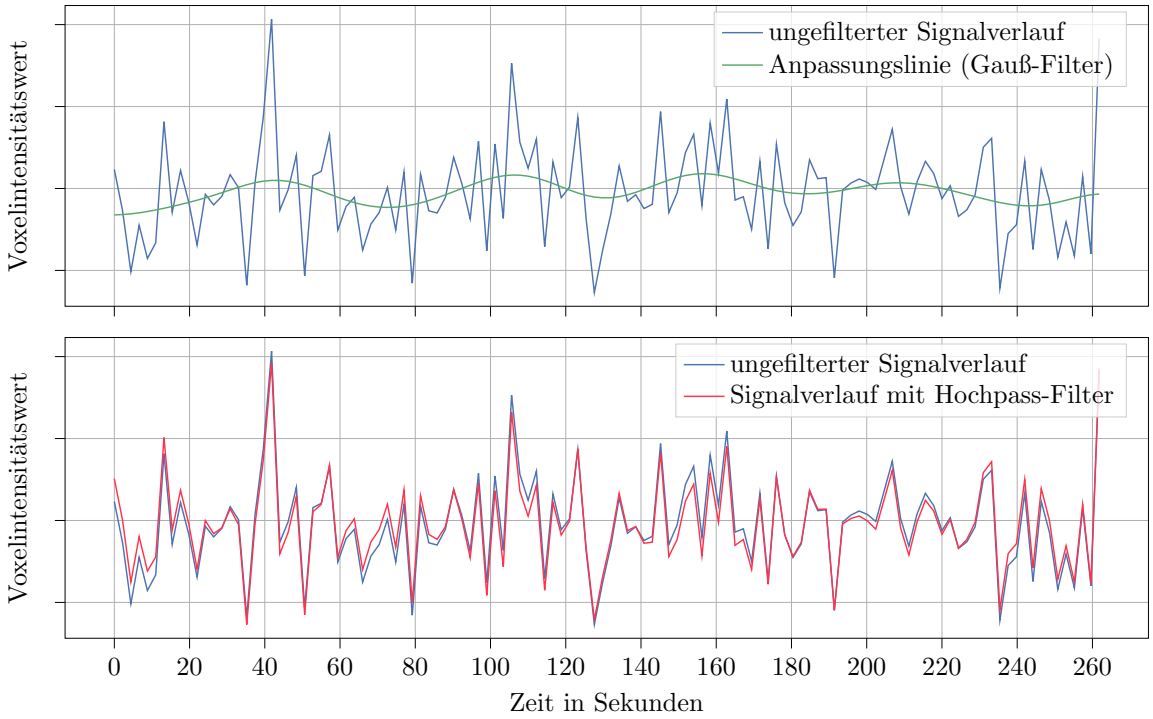

Abb. 2.8: **Zeitreihe eines Voxels vor und nach Anwendung des Hochpass-Filters.** Im ersten Schritt wird die Anpassungslinie (grün) durch Weichzeichnung des ungefilterten Signalverlaufs (blau) mit einem Gauß-Filter gebildet. Der mit einem Hochpass-Filter bereinigte Signalverlauf (rot) entsteht durch Subtraktion der Anpassungslinie vom ursprünglichen Signalverlauf.

und Grefkes, 2011). Diese zeichnen sich dadurch aus, dass der Proband – im Gegensatz zu *task-based*-Aufnahmen – für die Dauer der Bildakquisition gebeten wird, keine spezifische Aufgabe zu lösen, um sicherzustellen, dass sich das Gehirn in einem reizunabhängigen Zustand befindet. Die Hirnregionen, zwischen denen während dieses Ruhezustandes eine funktionelle Konnektivität besteht, werden als Ruhenetzwerke bezeichnet (Miall und Robertson, 2006). Eines der bekanntesten funktionellen Ruhenetzwerke ist das sog. *default mode network*, bei dem u. a. der mediale präfrontaler Kortex, der Praecuneus und Teile des Gyrus cinguli miteinander in Verbindung stehen (Andrews-Hanna et al., 2014). In den folgenden Abschnitten wird eine Methode zur Identifizierung dieser Netzwerke und deren Vergleich auf Probandenniveau mittels dualer Regression beschrieben.

2.4.1 Unabhängige Komponentenanalyse

Im Kontext von *neuroimaging* stellt die unabhängige Komponentenanalyse (*independent component analysis, ICA*) eine Methode zur Berechnung der zuvor beschriebenen räumlich unabhängigen und zeitlich synchronen Netzwerke dar. Dabei wird angenommen, dass eine funktionelle MRT-Aufnahme von mehreren, voneinander unabhängigen Signalen überlagert wird, die die neuronale Aktivität der Ruhenetzwerke, aber auch Bildrauschen repräsentieren, das z. B. durch Artefakte oder herzschlagsynchrone Bewegungen entstehen kann (Kiviniemi et al., 2003). Jede dieser voneinander unabhängigen Komponenten (*independent components, ICs*) besteht aus einer ‘räumlichen Karte’ (*spatial map*) sowie deren Zeitserie während der Bildakquisition. Dabei stellen die einzelnen Elemente der *spatial maps*, getrennt für jede Komponente, ein Maß für die Korrelationen eines Voxels mit den restlichen Voxeln der Aufnahme dar und somit auch für die Stärke der intrinsischen Konnektivität (Martuzzi et al., 2011). Die Zeitserie hingegen beschreibt die Verstärkung sowie die Unterdrückung des in der jeweiligen

spatial map festgehaltenen Signals im zeitlichen Verlauf (McKeown et al., 1998). Mathematisch kann eine funktionelle MRT-Aufnahme mit eine Matrix $\mathbf{M}_{\mathrm{jt}}$ beschrieben werden, wobei j der Anzahl der Voxel und t der Anzahl an *time points* entspricht. $\mathbf{M}_{\mathrm{jt}}$ setzt sich aus n verschiedenen *ICs* und einem weißen Rauschen $\mathbf{E}_{\mathrm{jt}}$ zusammen:

$$\mathbf{M}_{\mathrm{jt}} = \sum_{k=1}^{n} \mathbf{A}_{\mathrm{jk}} \mathbf{S}_{\mathrm{kt}} + \mathbf{E}_{\mathrm{jt}}$$

Die Spalten von $\mathbf{A}_{\mathrm{jk}}$ enthalten dabei die *spatial maps* der *ICs*, während die Zeilen von $\mathbf{S}_{\mathrm{kt}}$ deren jeweilige Zeitserie repräsentieren (McKeown et al., 2003). Je nachdem, ob nach räumlich oder zeitlich stochastisch unabhängigen Signalverläufen gesucht wird, sind die *spatial maps* oder die Zeitserien die unabhängige Variable der *ICA*, wobei im Fall einer funktionellen MRT-Aufnahme auf Grund der vergleichsweisen hohen Anzahl an Voxeln von einer räumlichen Unabhängigkeit ausgegangen werden kann (McKeown et al., 1998).

Zur Berechnung der *ICs* müssen die Matrizen $\mathbf{A}_{\mathrm{jk}}$ und $\mathbf{S}_{\mathrm{kt}}$ mit Hilfe komplexer Algorithmen in ihre Komponenten zerlegt werden. Populär sind dabei die Verfahren InfoMax (Bell und Sejnowski, 1995) und FastICA (Hyvarinen, 1999), die ausreichend gute Ergebnisse liefern, wenn bestimmte mathematische Annahmen über die Komponenten zugrunde gelegt werden (Daubechies et al., 2009) und somit auch für die Extraktion von Ruhenetzwerken geeignet sind. Die unterschiedlichen *ICs* lassen sich danach sortieren, wie viel Varianz der Aufnahme sie erklären, wobei zu beachten ist, dass bestimmte Komponenten wahrscheinlich nur Störsignale repräsentieren. Für einen Vergleich von Gruppen (z. B. Patienten mit IPS und GK) kann die *ICA* zudem auf Gruppenniveau durchgeführt werden, sodass für jede Gruppe eine *group IC-map* berechnet wird.

2.4.2 Duale Regression

Die duale Regression in FSL (Nickerson et al., 2017; Beckmann et al., 2009) ist eine Methode, die zum Vergleich der funktionellen Konnektivität zwischen verschiedenen Probanden verwendet werden kann, indem sie aus den *EPI-volumes* der Probanden die zuvor vorgegebene *spatial maps* mit ihren zugehörigen Zeitserien extrahiert. Die Vorgabe von *spatial maps* ermöglicht es, ausschließlich jene *ICs* der Probanden miteinander zu vergleichen, die für die jeweilige Fragestellung von Interesse sind und Komponenten auszuschließen, die beispielsweise auf Grund von Bildrauschen entstanden sind. Typischerweise entstammen die vorgegebenen *spatial maps* einer zuvor berechneten *group IC-map*, die somit zugleich die Anzahl und Art der *spatial maps* vorgibt, nach denen in den probandenspezifischen *EPI-volumes* gesucht werden soll.

Zu diesem Zweck wird in einem ersten Schritt die *group IC-map* in eine zweidimensionale Matrix $\hat{\mathbf{S}}$ der Form (Anzahl Voxel $\times$ Anzahl *ICs*) überführt. Entsprechend werden bei einer probandenspezifischen Aufnahme alle Voxel-Intensitäten aus den *EPI-volumes* mit ihren jeweiligen *time points* in eine Matrix $\mathbf{M}$ der Form (Anzahl Voxel $\times$ Anzahl *time points*) übertragen. Beide Matrizen stehen dabei in folgender Beziehung:

$$\mathbf{M} = \hat{\mathbf{S}}\hat{\mathbf{B}}_{\mathrm{TC}} + \mathbf{E}_1$$

$\hat{\mathbf{B}}_{TC}$ spiegelt die probandenspezifische Zeitserie jeder Komponente der *group IC-map* wider und $\mathbf{E}$ fasst mögliche Ungenauigkeiten, wie etwa Bildartefakte, in einer Fehlermatrix zusammen. Formal betrachtet hat $\hat{\mathbf{B}}_{TC}$ folglich die Form (Anzahl *ICs* $\times$ Anzahl *time points*) und kann mittels multivariate Regression berechnet werden. Nach dem Transponieren beider Matrizen wird in einem zweiten Schritt auf ähnliche Weise die probandenspezifische Zeitserie

verwendet, um anhand den korrespondierenden *EPI-volumes* $\mathbf{M}$ die probandenspezifischen *spatial maps* $\hat{\mathbf{B}}_{\mathrm{SM}}$ mit folgender Gleichung zu bestimmen:

$$\mathbf{M}^{\mathrm{T}} = (\hat{\mathbf{B}}_{\mathrm{TC}})^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{SM}} + \mathbf{E}_2$$

$\hat{\mathbf{B}}_{\mathrm{SM}}$ hat dabei die gewünschte Form (Anzahl *ICs* × Anzahl Voxel) und repräsentiert die *spatial maps* der probandenspezifischen Aufnahme mit den Komponenten aus der *group IC-map* (vgl. Abbildung 2.9). Um aussagekräftige Ergebnisse zu erhalten, ist es zwingend erforderlich, dass *group IC-map* und funktionelle MRT-Aufnahme in einem einheitlichen Standardraum liegen, damit Voxel mit gleichen Koordinaten auch gleiche Orte des Gehirns abbilden. Für Gruppenvergleiche von Patienten mit IPS und GK schlugen Griffanti et al. (2016) zudem vor, die *group IC-map* mit den Aufnahmen gesunder Probanden außerhalb der zu untersuchenden Kohorte zu erstellen.

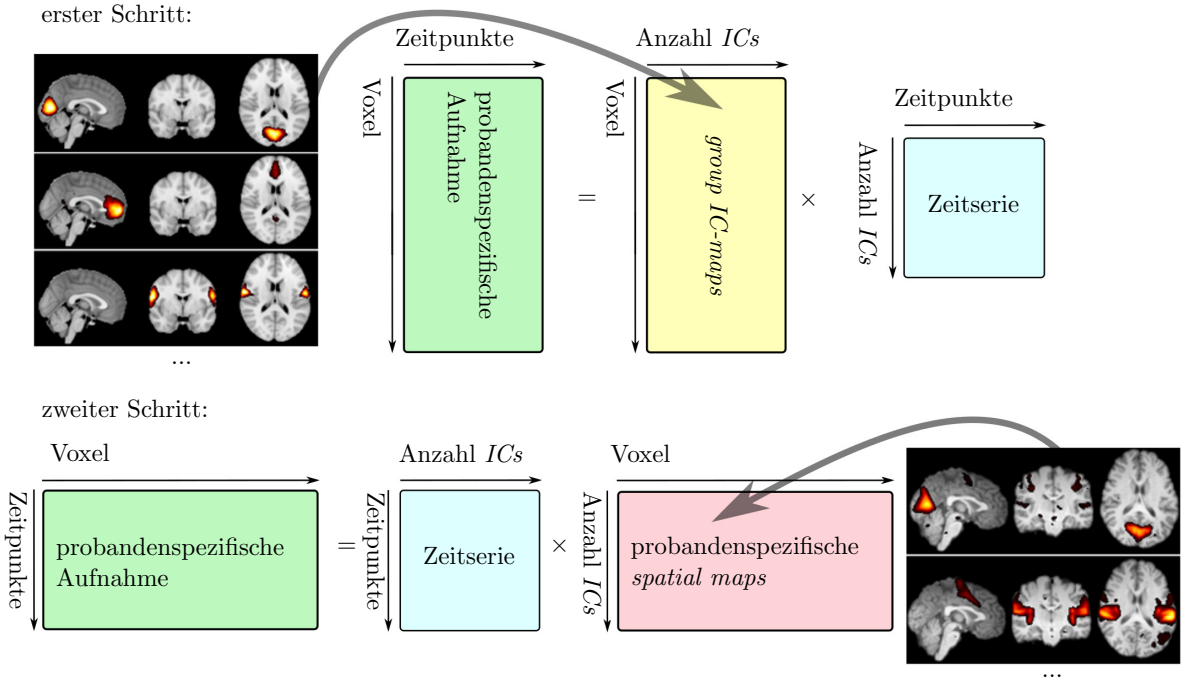

Abb. 2.9: **Prinzip der dualen Regression.** Im ersten Schritt werden die Zeitserien durch multivariate Regression berechnet. Diese werden im zweiten Schritt genutzt, um, ebenfalls durch multivariate Regression, die probandenspezifischen *spatial maps* aus den Aufnahmen zu extrahieren (erstellt in Anlehnung an Nickerson et al., 2017).

2.5 Studienspezifisches Matching

Störfaktoren (Confounder) sind externe Variablen, die eine scheinbare Beziehung zwischen den Eingabe- und Ausgabevariablen verzerren und somit zu fehlerhaften Schlussfolgerungen führen (Q. Zhao et al., 2020). In diesem Kontext beschreibt Matching ein Verfahren zur Minimierung des Einflusses von Confoundern in Studien (De Graaf et al., 2011). Dies wird dadurch erreicht, dass Probanden auf verschiedene Arten in Gruppen eingeteilt werden, die sich im Hinblick auf das zu untersuchende Merkmal unterscheiden, während die Confounder ausgeglichen in allen Gruppen verteilt sind. Nachfolgend wird ein Überblick über die Bedeutung von Confoundern im Kontext von *machine learning* gegeben und der Algorithmus vorgestellt, mit dem in der vorliegenden Studie das Matching durchgeführt wurde.

2.5.1 Confounder

Die Modelle zahlreicher *machine learning*-Verfahren können unerwünschte Störfaktoren aus den Trainingsdaten extrahieren und sie für ihre Vorhersage nutzen (Badgeley et al., 2019). Beispielsweise entdeckten Zech et al. (2018), dass bestimmte *CNNs* bei der Beurteilung von Röntgenaufnahmen des Thorax mit hoher Genauigkeit bestimmten, ob die zu analysierende Aufnahme in liegender oder stehender Position des Patienten erstellt wurde und diese Information verwendeten, um den Schweregrad einer Pneumonie zu klassifizieren. Das Modell folgerte nahezu unabhängig von der tatsächlichen Präsentation der Pneumonie, dass bei Liegendaufnahmen eine höhergradige Form der Erkrankung vorliegen müsse als bei einer Stehendaufnahme. Diese Annahme entstand vermutlich auf Grund des Fehlens von Stehendaufnahme von Patienten mit schwerer Pneumonie in den Trainingsdaten. Dies lag wiederum daran, dass schwer erkrankte Patienten nicht in der Lage waren, aufzustehen. Agniel et al. (2018) stellten in einer retrospektiven Beobachtungsstudie fest, dass ein Modell, das zeitliche Modalitäten von Laboruntersuchungen nutzte, wie beispielsweise den Wochentag und die Uhrzeit der Laborparametererhebung sowie den zeitlichen Abstand zwischen den Laboranforderungen, in vielen Fällen eine präzisere Vorhersage der Drei-Jahres-Mortalität ermöglichte als bei Verwendung der tatsächlich erhobenen Laborparameter. Die wahrscheinliche Erklärung dafür ist, dass bei schwerkranken Patienten häufiger Laboruntersuchungen angeordnet wurden als bei gesünderen. Um solche Effekte in der vorliegenden Studie zu minimieren, müssen entsprechende Confounder identifiziert und kontrolliert werden.

Ein wesentlicher Confounder bei der Analyse von MRT-Aufnahmen stellt generell das dabei verwendete Erfassungsprotokoll dar, das u. a. die Auswahl des Scannermodells und die Akquisitionsmodalitäten umfasst (Ferrari et al., 2020). Dieser Faktor kann bei der Analyse der HHU-Kohorte vernachlässigt werden, da alle Aufnahmen bei identischen Akquisitionsmodalitäten mit demselben MRT-Scanner erstellt wurden. Werden jedoch Aufnahmen aus unterschiedlichen Kohorten mit teils unterschiedlichen Erfassungsprotokollen verwendet, ist dieser Confounder problematisch und kann auch durch eine entsprechende Vorverarbeitung der Aufnahmen nur in einem begrenzten Umfang korrigiert werden.

Unter Verwendung von *deep learning*-Verfahren kann anhand von MRT-Aufnahmen des Gehirns sowohl das Alter (Huang et al., 2017) als auch das Geschlecht (Pawlowski und Glocker, 2019) eines Probanden bestimmt werden. Beide Parameter stellen bei den vorliegenden Bildgebungsmodalitäten folglich einen potentiellen Confounder dar (Kruggel et al., 2010; Wachinger et al., 2019), da sie in IPS- und GK-Gruppe nicht gleichmäßig verteilt sind. Demnach wäre eine Gruppenvorhersage anhand eines einzelnen Merkmals möglich, das in einer Gruppe überrepräsentiert ist, wie beispielsweise hohes Alter oder weibliches Geschlecht. Für die vorliegende Studie wurde daher ein Matching-Algorithmus entwickelt (vgl. Abschnitt 2.5.2), mit dem die Alters- und Geschlechterverteilung in IPS- und GK-Gruppe angeglichen werden kann.

2.5.2 Matching-Algorithmus

Der im Folgenden beschriebene Algorithmus ist dazu geeignet, die Verteilung eines stetigen und eines binären Merkmals innerhalb beliebig vieler Gruppen durch Verwerfung möglichst weniger Merkmalsträger soweit anzupassen, dass beide Merkmale innerhalb aller Gruppen fast beliebig gleich verteilt sind. In der vorliegenden Studie wurde dieser Algorithmus verwendet, um die Alters- und Geschlechterverteilung der Probanden aus IPS- und GK-Gruppe aneinander anzupassen.

Im ersten Schritt werden verschiedene Ausprägungen des stetigen Merkmals getrennt für

jede Gruppe in Klassen zusammengefasst. Dazu wird das Intervall, in dem sich alle Werte x dieses Merkmals befinden, in n Klassen $K_1, K_2, \ldots, K_n$ zerlegt, sodass gilt $K_j = (x_{j-1}; x_j]$ mit $j = 1, 2, \ldots, n$. Im nächsten Schritt werden innerhalb aller Gruppen die Merkmalsträger entsprechend dieser Klassen verteilt und es wird berechnet, wie groß der Anteil an Merkmalsträgern in einer Klasse im Vergleich zur Anzahl der Merkmalsträger in der gesamten Gruppe ist. Zur Identifizierung der Klasse K_j, in der sich der berechnete relative Anteil an Merkmalsträgern für alle Gruppen am stärksten unterscheidet, wird die Differenz dieser Anteile für die in den Gruppen korrespondierenden Klassen $K_1, K_2, \ldots, K_n$ gebildet und verglichen. Bei der gesuchten Klasse ist diese Differenz am größten, sodass ein Merkmalsträger aus dieser Klasse verworfen wird. Die Gruppe des betroffenen Merkmalsträgers ist jene mit dem größten relativen Anteil an Merkmalsträgern in der entsprechenden Klasse. Schließlich wird die Verteilung des binären Merkmals für alle Gruppen bestimmt und der Merkmalsträger ausgeschlossen, der die gesuchte Gruppe, die gesuchte Klasse und das für seine Gruppe überrepräsentierte binäre Merkmal aufweist. Für den Fall, dass kein Merkmalsträger die drei Anforderungen erfüllt, wird das binäre Merkmal für den Ausschluss nicht weiter berücksichtigt. Dieses Verfahren kann beliebig häufig wiederholt werden und führt nach einer Iteration zu einer höheren Homogenität der Gruppen im Hinblick auf mindestens eines der beiden Merkmale (vgl. Abbildung 2.10). Mittels Zweistichproben-t-Test, der auf einen signifikanten Unterschied zwischen beiden Gruppen im Hinblick auf das stetige Merkmal prüft, kann diese Angleichung schließlich quantifiziert werden.

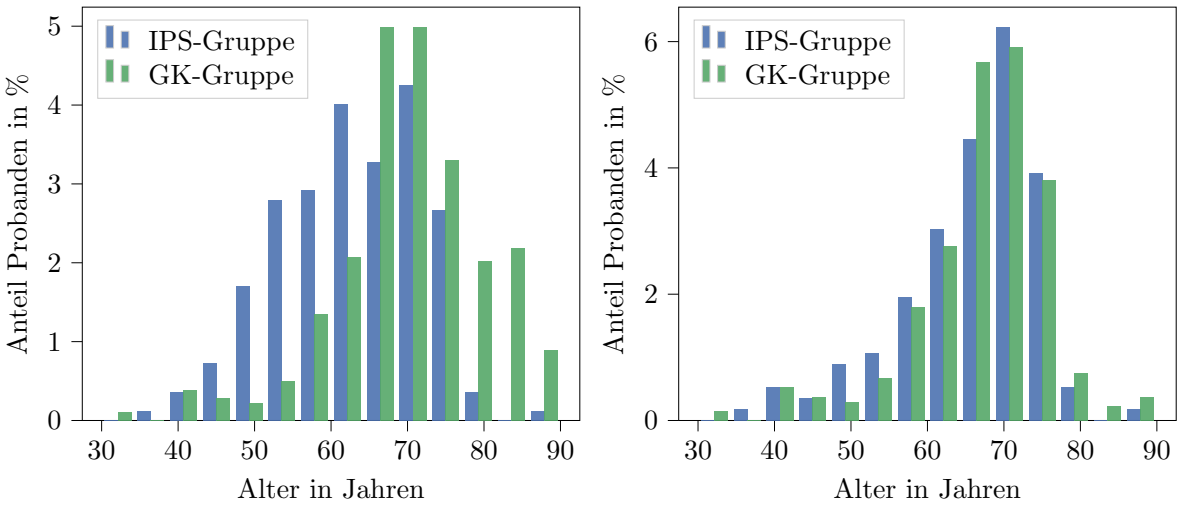

Abb. 2.10: **Beispielhafte Darstellung der Altersverteilung zweier Gruppen vor und nach Anwendung des Matching-Verfahrens.** Durch den Ausschluss von Probanden in der überrepräsentierten Altersklasse gleicht sich nach einigen Iterationen die Altersverteilung zwischen IPS- und GK-Gruppe an.

3 Material und Methoden

3.1 Datensammlung

Um die Vergleichbarkeit der Aufnahmen einer Bildgebungsmodalität untereinander zu verbessern, erfolgte deren Auswahl nach einheitlichen Ein- und Ausschlusskriterien, die in den Abschnitten zu den entsprechenden Kohorten erläutert sind (vgl. Abschnitt 3.1.1 und 3.1.2). In den externen Studien, in denen ein longitudinales Studiendesign verwendet wurde, sind zudem mehrere, im Abstand von wenigen Jahren erstellte Aufnahmen eines Probanden vorhanden. Die Verwendung dieser Aufnahmen als Trainingsdaten birgt das Risiko, dass eine Klassifizierung anhand individueller anatomischer Besonderheiten anstatt IPS-assoziierter Merkmale vorgenommen wird. Um diesen Effekt zu vermeiden, wurde je Proband nur die Aufnahme mit dem jüngsten Probandenalter zur weiteren Verarbeitung ausgewählt. Die Aufnahmen, die alle Kriterien erfüllten, wurden visuell auf Störfaktoren wie grobe Bildartefakte oder eine falsche Ausrichtung kontrolliert, um sie in solch einem Fall von der weiteren Verarbeitung auszuschließen. Für die Verwendung der im Folgenden beschriebenen Daten besteht ein positives Ethikvotum (Studiennummer 4878) der Ethikkommission an der Medizinischen Fakultät der HHU.

3.1.1 Interne Probandenkohorte

Die Datensätze der internen Kohorte ('HHU-Kohorte') entstammen einer an der HHU durchgeführten Bildgebungsstudie und beinhalten strukturelle T1-gewichtete, funktionelle und größtenteils auch diffusionsgewichtete MRT-Aufnahmen von 78 Patienten mit IPS und 78 GK (Mathys et al., 2016; Caspers et al., 2017, 2021). Die enthaltenen funktionellen Aufnahmen der Patienten während sog. Off-Phasen wurden in der vorliegenden Studie nicht berücksichtigt. Diese Phasen sind dadurch gekennzeichnet, dass die Wirkung der Antiparkinsonika nachlässt und betroffene Patienten verstärkt unter den typischen Symptomen leiden.

Auf Grund der unausgeglichenen Verteilung der potentiellen Confounder Alter und Geschlecht in IPS- und GK-Gruppe (vgl. Tabelle 3.1) wurde bei der HHU-Kohorte das in Abschnitt 2.5.2 beschriebene Matching-Verfahren angewandt. Für die Auswahl der zu verwerfenden Aufnahmen wurde zudem in Anlehnung an eine Studie von Huang et al. (2017) davon ausgegangen, dass der mittlere absolute Fehler bei der Altersbestimmung anhand von MRT-Aufnahmen des Gehirns vier Jahre beträgt. Daher kann das Alter innerhalb dieser Spannweite nicht weiter differenziert werden, sodass es ausreicht, für dieses Merkmal 15 Klassen zu bilden, um eine Spanne zwischen 30 und 90 Jahren abzubilden. Beurteilt wurde die Altersverteilung mit einem Zweistichproben-t-Test, der auf einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe prüft. Als Zielwert wurde $p > 0{,}1$ festgelegt. Die Geschlechterverteilung sollte als binäres Merkmal in IPS- und GK-Gruppe ebenfalls möglichst ausgeglichen sein und wurde mittels χ^2-Unabhängigkeitstest beurteilt. Auf Grund der bereits ausgewogenen Verteilung von Männern und Frauen in der ursprünglichen Kohorte war die Festlegung eines Zielwertes nicht erforderlich. Zum Erreichen beider Vorgabe mussten lediglich 21 der 156 Probanden ausgeschlossen werden. Die Charakteristika der IPS- und der GK-Gruppe vor und nach Matching sind in Tabelle 3.1 zusammengefasst.

Die Bildakquisition der strukturellen T1-gewichteten Aufnahmen erfolgte mit einem 3 T

MRT-Scanner des Modells Siemens Trio mit *MPRAGE*-Sequenz, *repetition time* = 2,3 s, *echo time* = 2,96 ms, *inversion time* = 90 ms, *flip angle* = 8 °, *field of view* = 240 mm × 256 mm *sagittal plane*, Anzahl Schichten = 192 und *voxel size* = 1 mm^3. Im gleichen Scanner wurden die rs-fMRT-Aufnahmen mit *EPI*-Sequenz, 300 *time points*, *repetition time* = 2,2 s, *echo time* = 30 ms, *flip angle* = 90 °, *field of view* = 200 mm × 200 mm *axial plane*, Anzahl Schichten = 36 und *voxel size* = 3,1 × 3,1 × 3,1 mm^3 bei einer Messzeit von ca. 11 min erstellt.

Zur internen Validierung aller für die vorliegende Studie erstellten Classifier wurden von der Kohorte nach Matching 20 % der Probanden zur Verwendung als Testdaten zurückgehalten. Deren Auswahl erfolgte in Form einer stratifizierten Randomisierung, um die Alters- und Geschlechterverteilung innerhalb der IPS- und GK-Gruppe in Trainings- und Testdaten mit größtmöglicher Ähnlichkeit abzubilden. Des Weiteren wurde festgelegt, dass die Testdaten Aufnahmen von 13 Patienten mit IPS und 14 GK enthalten sollen und zwischen ältestem und jüngstem Probanden in beiden Gruppen ein Altersunterschied von mindestens 40 Jahren bestehen muss. Durch diese Einschränkungen wurde eine hohe Heterogenität der Probanden in den Testdaten im Hinblick auf Gruppenzugehörigkeit, Alter und Geschlecht sichergestellt. Die daraus resultierende Zusammensetzung der entsprechenden Subkohorten ist in Tabelle 3.2 dargestellt. Weitere Merkmale der IPS-Gruppe nach Matching, wie beispielsweise die Stadieneinteilung nach Hoehn und Yahr (1967) sowie die Krankheitsdauer bis zur Bildakquisition, sind in Tabelle 3.4 angegeben.

	Original			nach Matching		
	n	davon Männer	Alter (Jahre)	n	davon Männer	Alter (Jahre)
IPS	78	50 (64 %)	61,4 ± 9,7	63	35 (56 %)	59,1 ± 9,5
GK	78	42 (54 %)	55,4 ± 10,6	72	41 (57 %)	56,4 ± 10,3
gesamt	156	92 (59 %)	58,4 ± 10,6	135	76 (56 %)	57,7 ± 10,0
p-Wert		0,25	< 0,001		0,99	0,12

Tabelle 3.1: **HHU-Kohorte vor und nach Matching.** Durch selektiven Ausschluss von 21 Probanden wurde die Alters- und Geschlechterverteilung in IPS- und GK-Gruppe aneinander angeglichen. Die p-Werte beziehen sich auf einen Zweistichproben-t-Test, der einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe untersucht bzw. auf einen χ^2-Unabhängigkeitstest für das Geschlecht in beiden Gruppen.

	Trainingsdaten			Testdaten		
	n	davon Männer	Alter (Jahre)	n	davon Männer	Alter (Jahre)
IPS	50	29 (58 %)	58,2 ± 8,8	13	6 (46 %)	62,7 ± 10,9
GK	58	34 (59 %)	56,2 ± 8,8	14	7 (50 %)	57,1 ± 14,8
gesamt	108	63 (58 %)	57,2 ± 8,9	27	13 (48 %)	59,8 ± 13,4
p-Wert		0,90	0,25		0,85	0,30

Tabelle 3.2: **Trainings- und Testdaten der HHU-Kohorte nach Matching.** Aus der Kohorte wurden nach Matching 20 % der Probanden als Testdaten abgespalten (*80:20 split*). Die p-Werte beziehen sich auf einen Zweistichproben-t-Test, der einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe untersucht bzw. auf einen χ^2-Unabhängigkeitstest für das Geschlecht in beiden Gruppen.

Da von einigen Probanden keine diffusionsgewichteten Aufnahmen akquiriert wurden, standen für den DTI-Classifier im Vergleich zu den anderen Classifiern weniger Trainings- und Testdaten zur Verfügung (vgl. Tabelle 3.3). Im Rahmen der 5-fachen Kreuzvalidierung war

diese Kohorte nach Matching ($n = 106$, davon 57 Patienten mit IPS und 49 GK) folglich ebenfalls anders als die zuvor beschriebene HHU-Kohorte nach Matching zusammengesetzt. Das durchschnittliche Alter der Patienten mit IPS betrug 59,5 ±9,4 Jahre und das der GK 55,5 ±9,6 Jahre. Der Anteil an Männern mit IPS lag bei 56 %, der von GK bei 61 %.

	Trainingsdaten			Testdaten		
	n	davon Männer	Alter (Jahre)	n	davon Männer	Alter (Jahre)
IPS	44	26 (59 %)	$58{,}6 \pm 8{,}7$	13	6 (46 %)	$62{,}7 \pm 10{,}9$
GK	40	26 (65 %)	$55{,}7 \pm 9{,}2$	9	4 (44 %)	$54{,}8 \pm 10{,}9$
gesamt	84	52 (62 %)	$57{,}2 \pm 9{,}1$	22	10 (45 %)	$59{,}5 \pm 11{,}6$
p-Wert		0,74	0,15		0,72	0,13

Tabelle 3.3: **Trainings- und Testdaten der HHU-Kohorte nach Matching mit diffusionsgewichteten Aufnahmen.** Im Unterschied zu Tabelle 3.2 wurden nur Probanden berücksichtigt, von denen zusätzlich diffusionsgewichteten Aufnahmen akquiriert wurden. Die p-Werte beziehen sich auf einen Zweistichproben-t-Test, der einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe untersucht bzw. auf einen χ^2-Unabhängigkeitstest für das Geschlecht in beiden Gruppen.

3.1.2 Externe Probandenkohorten

Unter anderem im Rahmen einer Studie von Mathys et al. (2016) wurden zur Untersuchung der funktionellen Konnektivität bei Patienten mit IPS, zusätzlich zur HHU-Kohorte, Datensätze an der Uniklinik RWTH Aachen ('Aachen-Kohorte', 22 Patienten mit IPS, 24 GK) und an der Uniklinik Köln ('Köln-Kohorte', 13 Patienten mit IPS, 13 GK) erhoben. Die für die rs-fMRT-Aufnahmen verwendeten Akquisitionsparameter (*EPI*-Sequenz, *repetition time* = 2,2 s, *echo time* = 30 ms und *voxel size* = $3{,}1 \times 3{,}1 \times 3{,}1\,\mathrm{mm}^3$) sind weitgehend identisch zu denen der HHU-Kohorte, jedoch wurde im Vergleich zu dieser eine geringere Anzahl an *volumes* akquiriert (vgl. Tabelle 3.5). Die Aufnahmen dieser Kohorten wurden in der vorliegenden Studie dazu genutzt, die externe Validität des rs-fMRT-Classifiers im Rahmen eines *leave-one-site-out*-Ansatzes (vgl. Abschnitt 3.4.2) zu überprüfen.

Die Parkinson's Progression Markers Initiative (PPMI, `ppmi-info.org`, Marek et al., 2011) ist eine externe, internationale und longitudinale Beobachtungsstudie, in deren Rahmen Datensätze von Patienten mit neu aufgetretenem und unbehandeltem IPS sowie von GK gesammelt werden. Einschlusskriterien für die IPS-Gruppe umfassen u. a. das Vorliegen mindestens zweier typischer Leitsymptome (vgl. Unterkapitel 1.1), Stadium I oder II nach Hoehn und Yahr zu Studienbeginn, der Nachweis eines Dopamintransportermangels und die Prognose, dass keine Antiparkinsonika bis mindestens sechs Monate nach Studienbeginn benötigt werden. Die Datensätze umfassen Messungen von unterschiedlichen bildgebenden Verfahren sowie genetische Untersuchungen, wobei für die vorliegende Studie ausschließlich strukturelle T1-gewichtete MRT-Aufnahmen verwendet wurden. Zur Auswahl dieser Aufnahmen ('PPMI-Kohorte', 192 Patienten mit IPS, 73 GK) wurden in der Datenbank folgende Sucheinstellungen angewandt: *modality*: MRI, *field strength*: 3 T, *weighting*: T1, *slice thickness*: 1 mm und *image description* mit 'MPRAGE'. Die übrigen Akquisitionparameter sind mit Ausnahme von *field of view* = 256 mm und Anzahl Schichten = 192 auf Grund der Verwendung unterschiedlicher MRT-Scanner uneinheitlich.

Zur Erhöhung der Anzahl an Aufnahmen von GK wurden diese um 73 Aufnahmen aus den Beobachtungsstudien der Alzheimer's Disease Neuroimaging Initiative ('ADNI', `adni.loni.usc.edu`, Jack Jr et al., 2008) sowie um 44 Aufnahmen aus dem Nathan Kline Institute-

Rockland Sample ('Rockland', `fcon_1000.projects.nitrc.org/indi/enhanced/`, Nooner et al., 2012) erweitert. Die erstgenannte Studie wurde im Jahr 2003 als öffentlich-private Partnerschaft initiiert und ihr Hauptziel bestand in der Überprüfung, ob MRT, Positronen-Emissions-Tomographie, andere biologische Marker sowie klinische und neuropsychologische Beurteilungen kombiniert werden können, um das Fortschreiten leichter kognitiver Beeinträchtigungen und der frühen Alzheimer-Krankheit zu messen. Ähnlich wie bei der PPMI umfasst diese Studie folglich neben struktureller T1-gewichteter MRT-Aufnahmen aus mehreren Zentren eine Vielzahl an weiteren Bildgebungsmodalitäten, die für die vorliegende Studie jedoch nicht berücksichtigt wurden. Die Auswahl der Aufnahmen erfolgte anhand identischer Sucheinstellungen wie für die PPMI-Kohorte, jedoch wurde die Obergrenze für das Probandenalter auf 90 Jahre festgelegt. Bei diesem Ausschlusskriterium wurde berücksichtigt, dass das Höchstalter der Probanden aus der PPMI-Studie bei 90 Jahren liegt und demnach ein Matching mit älteren Probanden aus anderen Studien nicht möglich gewesen wäre. Die Akquisitionparameter sind mit Ausnahme der Parameter für die Sucheinstellung (*modality*: MRI, *field strength*: 3 T, *weighting*: T1, *slice thickness*: 1 mm und *image description* mit 'MPRAGE') auf Grund der Verwendung unterschiedlicher MRT-Scanner erneut uneinheitlich. Dem Nathan Kline Institute-Rockland Sample liegt das Ziel zugrunde, eine phänotypisierte Stichprobe von Probanden im Alter zwischen 6 und 85 Jahren mittels *neuroimaging* und Genetik zu erstellen, wobei auch von dieser Kohorte für die vorliegende Studie ausschließlich strukturelle T1-gewichtete Aufnahmen verwendet wurden. Die Bildakquisition erfolgte dabei mit einem 3 T MRT-Scanner des Modells Siemens MAGNETOM Trio mit *MPRAGE*-Sequenz, *repetition time* = 2,5 s, *echo time* = 3,5 ms, *flip angle* = 8 °, *field of view* = 256 mm *transversal plane*, Anzahl Schichten = 192 und *voxel size* = $1 \times 1 \times 1\,\mathrm{mm}^3$.

Für die externe Validierung des rs-fMRT- sowie des multimodalen Classifiers wurden ebenfalls Datensätze einer öffentlich verfügbaren Studie verwendet, die die Aufnahmen der sog. NEUROCON-Kohorte (27 Patienten mit IPS, 16 GK) und die der sog. Tao Wu-Kohorte (20 Patienten mit IPS, 20 GK) umfasst (`fcon_1000.projects.nitrc.org/indi/retro/parkinsons.html`, Badea et al., 2017). Das Ziel dieser Studie war es, Veränderungen der funktionellen Konnektivität bei Patienten mit IPS innerhalb einer Kohorte zu finden und anschließend zu prüfen, ob diese Veränderungen auch in Aufnahmen externer Kohorten wiederzufinden sind. Die Bildakquisition der strukturellen T1-gewichteten Aufnahmen aus der NEUROCON-Kohorte erfolgte mit einem 1,5 T MRT-Scanner des Modells Siemens Avanto mit *MPRAGE*-Sequenz, *repetition time* = 1,94 s, *echo time* = 3,08 s, *voxel size* = $0{,}97 \times 0{,}97 \times 1\,\mathrm{mm}^3$. Im Unterschied dazu erfolgte die Bildakquisition der strukturellen T1-gewichteten Aufnahmen aus der Tao Wu-Kohort mit einem 3 T MRT-Scanner des Modells Siemens Trio mit *MPRAGE*-Sequenz, *repetition time* = 1,1 s, *echo time* = 3,39 s, *voxel size* = $1 \times 1 \times 1\,\mathrm{mm}^3$. Die Alters- und Geschlechterverteilung, das Stadium nach Hoehn und Yahr sowie die Krankheitsdauer bis zur Bildakquisition sind in Tabelle 3.4 angegeben, die Parameter für die Akquisition der verwendeten rs-fMRT-Aufnahmen sind in Tabelle 3.5 zusammengefasst.

3.2 Uni- und multimodale Modelle auf interner Kohorte

Die Datensätze für Training und Validierung der nachfolgend beschriebenen Classifier stammten ausschließlich von den Probanden der HHU-Kohorte nach Matching. Dieses Vorgehen dient daher dazu, die Classifier zu entwickeln und deren interne Validität zu überprüfen.

3.2.1 T1-Classifier

Der T1-Classifier nutzte nach umfangreicher Vorverarbeitung ein *CNN* zur Klassifizierung struktureller T1-gewichteter MRT-Aufnahmen (vgl. Abbildung 3.1). Mit SPM12 wurde dazu

	n	davon Männer	Alter (Jahre)	Hoehn und Yahr	Krankheitsdauer (Jahre)
HHU					
IPS	63	35 (56 %)	59,1 ± 9,5	2,5 (2,0 – 3,0)[1]	7,8 ± 4,5[2]
GK	72	41 (57 %)	56,4 ± 10,3		
p-Wert		0,99	0,12		
Aachen					
IPS	22	14 (64 %)	65,3 ± 8,6	1,0 (1,0 – 2,0)	3,8 ± 3,3
GK	24	14 (58 %)	63,2 ± 4,9		
p-Wert		0,95	0,32		
Köln					
IPS	13	13 (100 %)	63,5 ± 7,6	2,0 (2,0 – 2,0)[3]	6,2 ± 2,9
GK	13	13 (100 %)	62,2 ± 5,8		
p-Wert		N/A	0,64		
NEUROCON					
IPS	27	17 (63 %)	68,7 ± 10,4	2,0 (2,0 – 2,0)	N/A
GK	16	4 (25 %)	67,6 ± 11,5		
p-Wert		0,04	0,76		
Tao Wu					
IPS	20	11 (55 %)	65,2 ± 4,3	2,0 (1,0 – 2,5)	5,4 ± 3,8
GK	20	12 (60 %)	64,8 ± 5,4		
p-Wert		1,0	0,78		

Tabelle 3.4: **Merkmale der unterschiedlichen Kohorten.** Die p-Werte beziehen sich auf einen Zweistichproben-t-Test, der einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe untersucht bzw. auf einen χ^2-Unabhängigkeitstest für das Geschlecht in beiden Gruppen. Für Stadium nach Hoehn und Yahr sind der Median sowie der Interquartilsabstand angegeben.
[1] Die Angaben waren bei 58 Patienten mit IPS vorhanden.
[2] Die Angaben waren bei 62 Patienten mit IPS vorhanden.
[3] Die Angaben waren bei 8 Patienten mit IPS vorhanden.

eine automatische Nullpunktkorrektur durchgeführt, die den Koordinatenursprung aller Aufnahmen auf die Commissura anterior festlegte. Anschließend wurde mit CAT12 das Verfahren der Voxel-basierten Morphometrie (vgl. Abschnitt 2.3.4) unter Verwendung der von der Software voreingestellten Standardparameter angewandt. Dabei wurden die Aufnahmen mittels Bildregistrierung bzw. Normalisierung in einen Standardraum mit 1,5 mm isometrischer Voxelgröße überführt, nach Gewebetyp segmentiert und moduliert, woraufhin der Anteil an grauer Substanz, weißer Substanz und Liquor cerebrospinalis in jedem Voxel der Aufnahme abgeschätzt wurde. Die weitere Analyse erfolgte mit den parametrischen Karten, in denen ausschließlich die graue Substanz zur Darstellung kommt (*'mwp1*' files*). Dieser Schritt ist damit zu begründen, dass Veränderungen der grauen Substanz bei Patienten mit IPS wesentlich häufiger beschrieben wurden als Veränderungen der weißen Substanz (Sarasso et al., 2021).

Die Architektur des verwendeten *CNN*-Modells ist in Tabelle 3.6 dargestellt, als Optimierungsalgorithmus wurde *ADAM* bei einer *learning rate* von 0,0001 verwendet. Der grundlegende Aufbau des Modells stammt aus der visuellen Klassifizierung zweidimensionaler Bilder (u. a. Q. Li et al., 2014; Sultana et al., 2018) und wurde für die vorliegende Studie soweit angepasst und optimiert, dass dieser zur Verarbeitung komplexer dreidimensionaler Eingabedaten eingesetzt werden kann. Nach 100 Trainingsepochen mit den Trainingsdaten ($n = 108$) der HHU-Kohorte nach Matching wurde das Training beendet, da sich darüber hinaus keine signifikante Verbesserung in Bezug auf die Modellgeneralisierbarkeit zeigte. Die Modellevaluation erfolgte mittels *hold-out validation* im *80:20 split* durch Validierung anhand der zurückgehaltenen Testdaten ($n = 27$). Damit alle verfügbaren Datensätze aus der Kohor-

Kohorte	MRT-Scanner	*TR* (s)	*TE* (ms)	*voxel size* (mm^3)	Anzahl *volumes*
HHU	Siemens Trio 3 T	2,2	30	$3{,}1 \times 3{,}1 \times 3{,}1$	300
Aachen	Siemens Trio 3 T	2,2	30	$3{,}1 \times 3{,}1 \times 3{,}1$	270
Köln	Siemens Trio 3 T	2,2	30	$3{,}1 \times 3{,}1 \times 3{,}1$	183
NEUROCON	Siemens Avanto 1,5 T	3,48	50	$3{,}8 \times 3{,}8 \times 5$	137
Tao Wu	Siemens Trio 3 T	2,0	40	$4 \times 4 \times 5$	239

Tabelle 3.5: **Scanner und Akquisitionsparameter für die Erstellung der rs-fMRT-Aufnahmen.** Die Aufnahmen wurden jeweils mit einer *EPI*-Sequenz akquiriert.

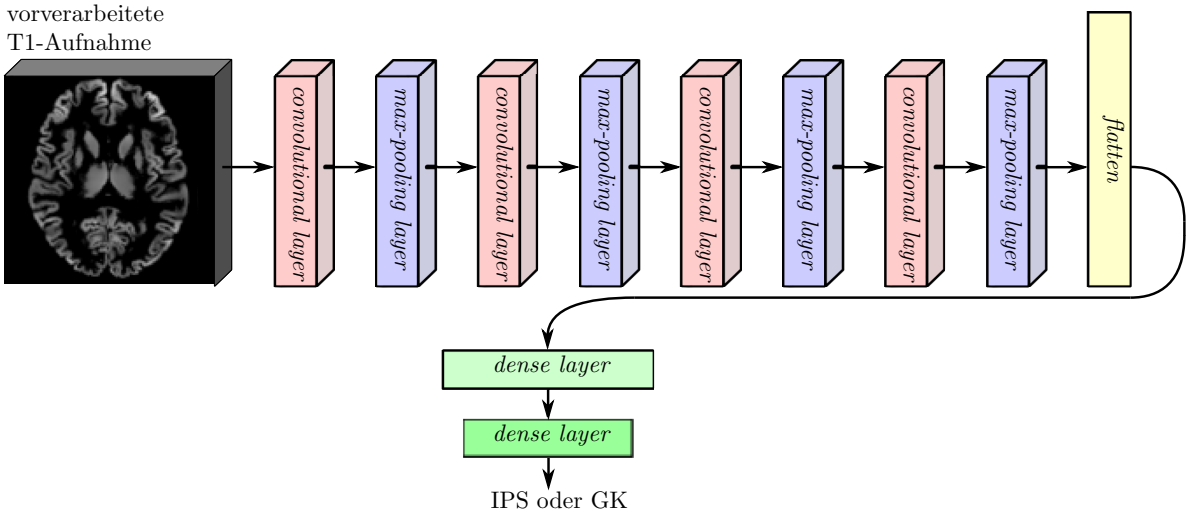

Abb. 3.1: **Pipeline des T1-Classifiers.** Nach umfangreicher Vorverarbeitung wurden die strukturellen T1-gewichteten Aufnahmen mit Hilfe eines dreidimensionalen *CNN* klassifiziert.

te nach Matching ($n = 135$) sowohl zum Training als auch zur Validierung genutzt werden konnten, wurden diese zudem für eine Modellevaluation mittels 5-facher Kreuzvalidierung verwendet (vgl. Abschnitt 2.1.5). Da bei jedem Trainingsdurchlauf die initialen Gewichte des Modells (*weights* und *biases*) zufällig festgelegt werden, wurden sowohl für die *hold-out validation* im *80:20 split* als auch für die 5-fache Kreuzvalidierung alle Trainingsdurchläufe insgesamt zehnmal durchgeführt.

Zur visuellen Hervorhebung der Bereiche einer Aufnahme, die zu einer korrekten Vorhersage des *CNN* geführt haben, wurden *heatmaps* erstellt. Die Erzeugung dieser *heatmaps* basierte auf der *gradient-weighted class activation mapping*-Methode von Selvaraju et al. (2017) und wurden bei den Patienten mit IPS angewandt, bei denen auch eine hohe Wahrscheinlichkeit für das Vorliegen von IPS berechnet wurde. Da der ursprüngliche Algorithmus der Autoren lediglich für zweidimensionale Bilder veröffentlicht wurde, musste er in vorliegender Studie soweit angepasst werden, dass er auch dreidimensionale Aufnahmen verarbeiten konnte. Als weitere Anpassung wurden die aus dem vorletzten *convolutional layer* gewonnenen Informationen zur Erstellung der *heatmaps* verwendet, da diese Schicht im Vergleich zum Vorschlag der Autoren, das letzte *convolutional layer* zu verwenden, auf Grund der höheren Auflösung den wahrscheinlich besseren Kompromiss zwischen hoher Bildsemantik und detaillierten räumlichen Informationen bietet. Als Konsequenz dieses Schrittes können die *features*, die zu einer bestimmten Vorhersage führten, präziser auf der ursprünglichen Aufnahme lokalisiert werden, sind dafür jedoch weniger komplex. Zur Visualisierung der berechneten

layer	*kernel size*	*output shape*	Aktivierungsfunktion
Eingabedaten		$91 \times 133 \times 92 \times 1$	
convolutional layer	$3 \times 3 \times 3$	$89 \times 131 \times 90 \times 16$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$44 \times 65 \times 45 \times 16$	
convolutional layer	$3 \times 3 \times 3$	$42 \times 63 \times 43 \times 32$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$21 \times 31 \times 21 \times 32$	
convolutional layer	$2 \times 2 \times 2$	$20 \times 30 \times 20 \times 64$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$10 \times 15 \times 10 \times 64$	
convolutional layer	$2 \times 2 \times 2$	$9 \times 14 \times 9 \times 128$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$4 \times 7 \times 4 \times 128$	
dropout 50 %		$6 \times 7 \times 6 \times 128$	
flatten		14336	
dense layer		128	
dropout 50 %		128	
dense layer		1	Sigmoidfunktion

Tabelle 3.6: **Architektur des im T1-Classifier verwendeten *CNN*-Modells.**

dreidimensionalen *heatmaps* wurden diese mit der Software NiBabel (Brett et al., 2020) zu NIfTI-Dateien konvertiert und in FSLeyes (McCarthy, 2020) mit der korrespondierenden Aufnahme überlagert. Die zur Klassifizierung führenden hirnstrukturellen Veränderungen eines Patienten mit IPS werden somit in seiner Aufnahme farblich hervorgehoben.

3.2.2 rs-fMRT-Classifier

Für den rs-fMRT-Classifier wurde ein *LSTM*-Modell herangezogen, das die aus den rs-fMRT-Aufnahmen extrahierten Zeitserien zur Klassifizierung nutzte (vgl. Abbildung 3.2). Relevante räumliche Informationen aus den Aufnahmen – und somit auch deren *spatial maps* – sind dadurch in neurobiologisch sinnvoller Art und Weise konserviert und müssen folglich nicht weiter modelliert werden. Die Vorverarbeitung aller Aufnahmen umfasste die Bewegungskorrektur der *EPI-volumes* mit FSL MCFLIRT sowie die Bildregistrierung in den MNI152-Standardraum mit 2 mm isometrischer Voxelgröße anhand korrespondierender hochauflösender T1-gewichteter Aufnahmen mittels FSL FEAT, FLIRT und FNIRT (Andersson et al., 2007). Darüber hinaus wurden *brain extraction* mit FSL BET, *spatial smoothing* durch einen Gauß-Filter mit *FWHM* von 5 mm, *grand mean scaling* mit $k = 10.000$ sowie *highpass temporal filtering* durch Subtraktion einer mit sigma = 150 s weichgezeichneten Anpassungslinie gemäß der in Abschnitt 2.3.8 beschriebenen Methode realisiert.

Anschließend folgte die Berechnung der probandenspezifischen Zeitserien mittels duale Regression der vorverarbeiteten *EPI-volumes* mit zwei unterschiedlichen *group IC-maps*. Die erste *group IC-map* umfasste die *spatial maps* von 100 *ICs*, die anhand der vorverarbeiteten Aufnahmen 812 gesunder Probanden aus der Datenbank des Human Connectome Project (Van Essen et al., 2013) berechnet wurden. Sie wurde im Rahmen der genannten Studie erstellt und im '1200 Subjects Data Release' (`humanconnectome.org/storage/app/media/documentation/s1200/HCP1200-DenseConnectome+PTN+Appendix-July2017.pdf`, WU-Minn HCP, 2017) veröffentlicht. Dieses Vorgehen entspricht der Empfehlung von Griffanti et al.

(2016), für Gruppenvergleiche zwischen Patienten mit IPS und GK eine *group ICA* mit Aufnahmen gesunder Probanden außerhalb der zu untersuchenden Kohorte durchzuführen und die daraus entstandene *group IC-map* zur Erkennung von krankheitsbedingten Veränderungen der Konnektivität innerhalb der zu untersuchenden Kohorte zu verwenden. Die Entscheidung, 100 *ICs* zu untersuchen, beruht auf einer Studie von Rubbert et al. (2019), die bei Verwendung dieser Granularität gute Ergebnisse für die Differenzierung zwischen Patienten mit IPS und GK anhand von Aufnahmen aus der HHU-Kohorte erzielten. Zusätzlich wurde eine eigene *pseudo group IC-map* zur Negativkontrolle angefertigt. Diese bestand aus den *spatial maps* von 100 pseudo-*ICs*, die lediglich ein über alle Hirnregionen verteiltes weichgezeichnetes Rauschen beinhalteten (vgl. Abbildung 3.3).

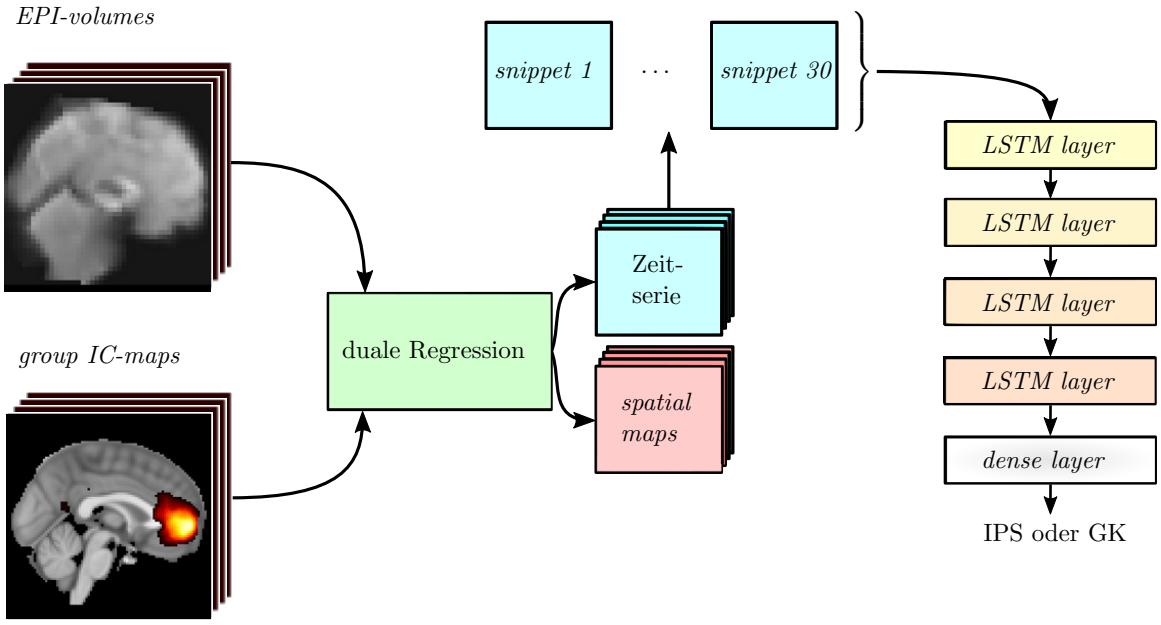

Abb. 3.2: **Pipeline des rs-fMRT-Classifiers.** Nach der dualer Regression wurde die berechneten Zeitserie in 30 *snippets* aufgeteilt, die mit Hilfe eines *LSTM*-Modells klassifiziert wurden. Für die *spatial maps* bestand hingegen keine weitere Verwendung.

Schließlich wurden alle Zeitserien bei einer Länge von 300 *time points* entlang der Zeitachse in jeweils 30 'Schnipsel' (*snippets*) mit gleicher Länge von jeweils 10 *time points* aufgeteilt, wobei ein *snippet* einem Datensatz entsprach und eine Zeitspanne von 22 s abdeckte. Zur Klassifizierung der *snippets* wurde ein *LSTM*-Modell mit der in Tabelle 3.7 gezeigten Architektur und *ADAM* als Optimierungsalgorithmus bei einer *learning rate* von 0,0001 verwendet. Nach 40 Trainingsepochen mit den Trainingsdaten ($n = 108$, entsprechend 3240 *snippets*) der HHU-Kohorte nach Matching wurde das Training beendet, da sich darüber hinaus keine signifikante Verbesserung in Bezug auf die Modellgeneralisierbarkeit zeigte. Die Modellevaluation erfolgte analog zum T1-Classifier mittels *hold-out validation* im *80:20 split* durch Validierung anhand der zurückgehaltenen Testdaten ($n = 27$, entsprechend 810 *snippets*). Um alle verfügbaren Datensätze sowohl für das Training als auch zur Validierung nutzen zu können, wurde die Modellevaluation zudem mittels 5-facher Kreuzvalidierung auf Probandenniveau an der gesamten Kohorte nach Matching ($n = 135$, entsprechend 4050 *snippets*) durchgeführt.

layer	*units*	*dropout*	*output shape*	Aktivierungsfunktion
Eingabedaten			10×100	
LSTM layer	40	20 %	10×40	hyperbolischer Tangens
LSTM layer	30	20 %	10×30	hyperbolischer Tangens
LSTM layer	20	20 %	10×20	hyperbolischer Tangens
LSTM layer	10	20 %	10	hyperbolischer Tangens
dense layer			2	Sigmoidfunktion

Tabelle 3.7: **Architektur des im rs-fMRT-Classifier verwendeten *LSTM*-Modells am Beispiel einer *IC-map* mit 100 Komponenten.** Bei einer anderen Anzahl an Komponenten ändert sich *output shape* entsprechend.

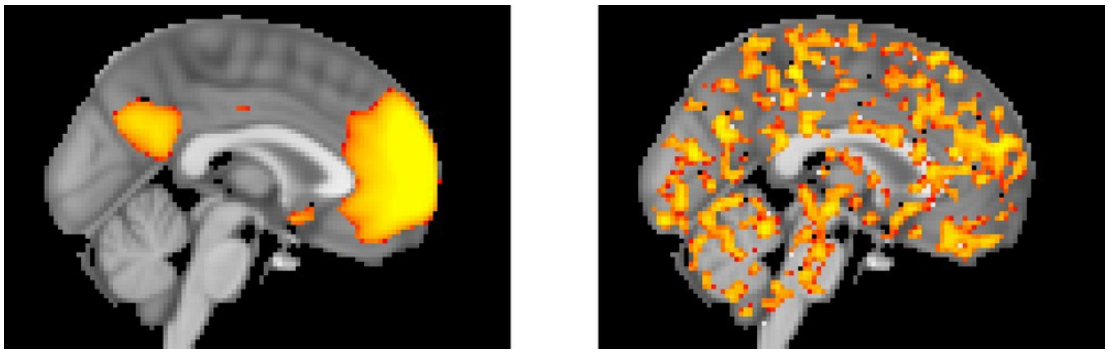

Abb. 3.3: **Vergleich einer echten *group IC-map* mit einer *pseudo group IC-map*.** Die echte *group IC-map* (links) wurde mittels *group ICA* mit Aufnahmen aus der Datenbank des Human Connectome Project berechnet, während die *pseudo group IC-map* (rechts) lediglich weichgezeichnetes Rauschen enthält. In der echten *group IC-map* repräsentieren die farbigen Bereiche zeitlich korrelierte Schwankungen der Voxel-Intensitäten, die somit auf eine mögliche funktionelle Konnektivität dieser Regionen hindeuten. Eine rötliche Färbung entspricht dabei einer schwachen Korrelation, während eine gelbliche Färbung auf eine starke Korrelation hinweist.

3.2.3 DTI-Classifier

Der DTI-Classifier basierte auf einem *CNN* mit ähnlicher Architektur wie das *CNN* des T1-Classifiers, um damit vorverarbeitete, diffusionsgewichtete MRT-Aufnahmen zu klassifizieren (vgl. Abbildung 3.4). Die Vorverarbeitung der Aufnahmen umfasste *brain extraction* unter Verwendung von FSL BET, eine Bewegungskorrektur sowie die Berichtigung wirbelstrominduzierter Verzerrungen (*eddy currents*) mittels FSL eddy (Andersson und Sotiropoulos, 2016). Nach Bestimmung der Diffusionstensoren mit Hilfe von FSL dtifit erfolgte daraus die Berechnung der fraktionalen Anisotropie, der radialen Diffusionsfähigkeit und der mittleren sowie der axialen Diffusivität. Schließlich wurden mittels FSL FLIRT die berechneten *volumes* in den MNI152-Standardraum mit 1 mm isometrischer Voxelgröße überführt.

Um alle berechneten *DTI*-Metriken zu berücksichtigen, wurden diese getrennt über jeweils einen Eingangskanal (*channel*) an das *CNN* übergeben. Bei diesem Schritt handelt es sich um ein neuartiges Verfahren, da die *channels* üblicherweise dazu verwendet werden, separat die Rot-, die Grün- und die Blauanteile einer Bilddatei für die Klassifizierung zu verarbeiten. In Tabelle 3.8 ist die Architektur des *CNN*-Modells dargestellt. Der dabei verwendete Optimierungsalgorithmus war *ADAM* bei einer *learning rate* von 0,0001. Nach 35

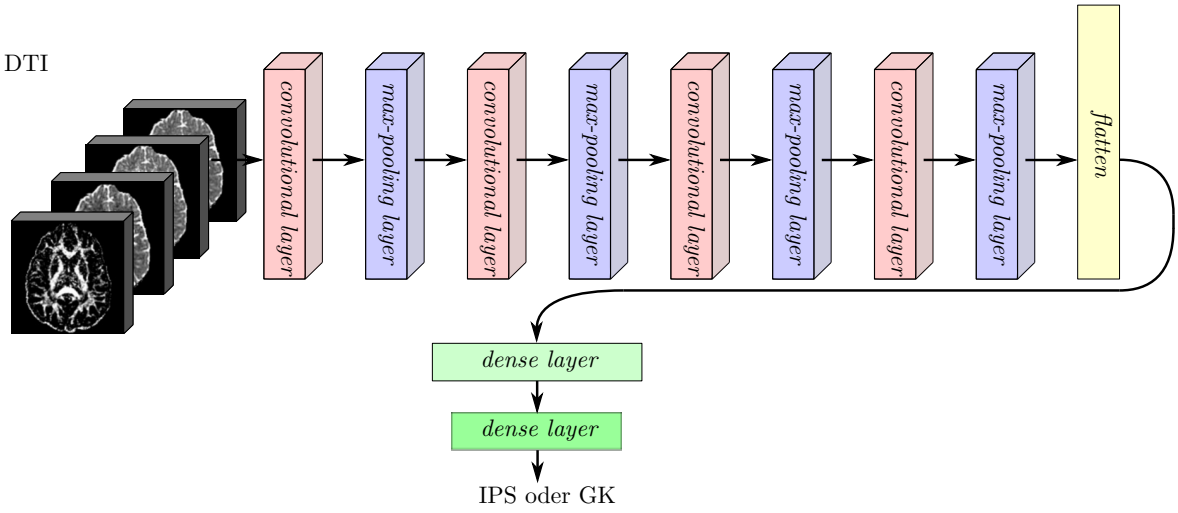

Abb. 3.4: **Pipeline des DTI-Classifiers.** Die *DTI*-Metriken fraktionale Anisotropie, radiale Diffusionsfähigkeit, axiale sowie mittlere Diffusivität einer Aufnahme wurden für Training und Validierung separat voneinander an ein dreidimensionales *CNN* übergeben und mit Hilfe dessen klassifiziert.

Trainingsepochen mit den Trainingsdaten ($n = 84$) der HHU-Kohorte nach Matching wurde das Training beendet, da sich darüber hinaus keine signifikante Verbesserung in Bezug auf die Modellgeneralisierbarkeit zeigte. Die Modellevaluation erfolgte analog zum T1-Classifier mittels *hold-out validation* im *80:20 split* durch Validierung anhand der zurückgehaltenen Testdaten ($n = 22$). Um alle verfügbaren Datensätze sowohl für das Training als auch zur Validierung nutzen zu können, wurde die Modellevaluation zudem mittels 5-facher Kreuzvalidierung an der gesamten Kohorte nach Matching ($n = 106$) durchgeführt.

layer	*kernel size*	*output shape*	Aktivierungsfunktion
Eingabedaten		$122 \times 173 \times 150 \times 4$	
convolutional layer	$3 \times 3 \times 3$	$120 \times 171 \times 148 \times 16$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$60 \times 85 \times 74 \times 16$	
convolutional layer	$3 \times 3 \times 3$	$58 \times 83 \times 72 \times 32$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$29 \times 41 \times 36 \times 32$	
convolutional layer	$2 \times 2 \times 2$	$28 \times 40 \times 35 \times 64$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$14 \times 20 \times 17 \times 64$	
convolutional layer	$2 \times 2 \times 2$	$13 \times 19 \times 16 \times 128$	*ReLU*
max-pooling layer	$2 \times 2 \times 2$	$6 \times 9 \times 8 \times 128$	
dropout 50 %		$6 \times 9 \times 8 \times 128$	
flatten		55296	
dense layer		128	
dropout 50 %		128	
dense layer		1	Sigmoidfunktion

Tabelle 3.8: **Architektur des im DTI-Classifier verwendeten *CNN*-Modells.**

3.2.4 Multimodaler Classifier

Der multimodale Classifier wurde mit dem Ziel konzipiert, eine finale Klassifizierung als Patient mit IPS oder GK auf Grundlage der teilweise für jede Bildgebungsmodalität unterschiedlichen Klassifizierungsergebnisse eines Probanden unter Berücksichtigung von Alter und Geschlecht zu berechnen. Als Eingabedaten verwendete dieser Classifier folglich neben Angaben zum Alter und Geschlecht die Klassifizierungsergebnisse der unimodalen Classifier in Form von Gleitkommazahlen zwischen 0 und 1. Bei Probanden ohne diffusionsgewichteten Aufnahmen wurde für das fehlende Klassifizierungsergebnis dieser Bildgebungsmodalität ein Wert von 0,5 angenommen, der folglich genau in der Mitte des Intervalls aller Klassifizierungsergebnisse liegt. Das Alter wurde unverändert übernommen, das Geschlecht als 0 oder 1 codiert. Mit diesen Eingabedaten wurden sowohl ein Modell mittels *logistic regression* (*LR*, Cox, 1958) als *machine learning*-Algorithmus unter Verwendung der voreingestellten Hyperparameter als auch zum Vergleich ein weiteres Modell mit *support vector machine* (*SVM*, Cortes und Vapnik, 1995) unter Einsatz eines linearen Kernels trainiert. Andere klassische *machine learning*-Algorithmen, wie *k-nearest neighbor*, *random forest* oder *XGboost*, wurden ebenfalls an den Datensätzen getestet, jedoch zeigte sich, dass diese den verwendeten Verfahren unterlegen waren.

Darüber hinaus wurde der multimodale Classifier mit zwei alternativen Varianten an Eingangsdaten trainiert und anschließend validiert. Dies diente der Analyse, inwiefern sich seine Vorhersagegenauigkeit bei einer Reduzierung der Eingabedaten verändert. Das zuvor beschriebene Vorgehen, bei dem die Klassifizierungsergebnisse aller Bildgebungsmodalitäten in Verbindung mit Angaben zum Alter und Geschlecht der Probanden verwendet wurden, diente als Referenzverfahren. In Variante 1 wurden die Angaben zum Alter und Geschlecht nicht berücksichtigt. In Variante 2 wurde der multimodale Classifier ausschließlich mit den Klassifizierungsergebnissen des rs-fMRT-Classifiers in Kombination mit Angaben zum Alter und Geschlecht der Probanden trainiert. Der rs-fMRT-Classifier wurde auf Grund seiner im Vergleich zu den anderen beiden unimodalen Classifiern höheren Vorhersagegenauigkeit ausgewählt.

Für die Validierung mittels *hold-out validation* im *80:20 split* anhand der Testdaten ($n = 27$) mussten zuerst die Eingabedaten für das Training in Form von Klassifizierungsergebnissen der unimodalen Classifier berechnet werden. Die Aufnahmen aus der HHU-Kohorte nach Matching wurden dazu getrennt für jede Bildgebungsmodalität verwendet, um mit dem T1-, dem rs-fMRT- bzw. dem DTI-Classifier durch 10-fache Kreuzvalidierung die Klassifizierungsergebnisse für alle Probanden aus den Trainingsdaten ($n = 108$) zu erzeugen. Die Eingabedaten der Testdaten entsprachen neben Angaben zum Alter und Geschlecht den Klassifizierungsergebnissen der unimodalen Classifier, die im Rahmen derer *hold-out validation* im *80:20 split* bereits berechnet worden waren. Für die 5-fache Kreuzvalidierung entsprachen die Eingabedaten der gesamten Kohorte nach Matching ($n = 135$) neben Angaben zum Alter und Geschlecht den Klassifizierungsergebnissen der unimodalen Classifier aus der unimodalen 5-fachen Kreuzvalidierung. Die verwendete Partitionierung der Probanden erfolgte zufällig und unabhängig von den Partitionierungen, die zuvor zur Evaluation der unimodalen Classifier verwendet worden waren. Sowohl für die *hold-out validation* im *80:20 split* als auch für die 5-fache Kreuzvalidierung wurden insgesamt zehn Trainingsdurchläufe durchgeführt, sodass alle Klassifizierungsergebnisse aus den zehn vorherigen Trainingsdurchläufen der unimodalen Classifier als Eingabedaten verwendet werden konnten.

Zur Präzisierung der Vorhersagewerte des multimodalen Classifier für das Vorliegen von IPS wurde eine Modellkalibrierung durchgeführt (vgl. Abschnitt 2.1.5). Dazu wurden ausgehend von den im Rahmen der 5-fachen Kreuzvalidierung berechneten Klassifizierungsergebnissen

für die gesamte Kohorte nach Matching ($n = 135$) die Datensätze 40 zufällig ausgewählter Probanden als Testdaten abgespalten und die Datensätze der übrigen Probanden ($n = 95$) als Trainingsdaten verwendet. Die Berechnung der Kalibrierungskurven erfolgte auf Basis dieser Trainingsdaten und unter Anwendung von *LR* bzw. *SVM* zur Bestimmung der vorhergesagten Wahrscheinlichkeit für IPS (vgl. Abbildung 3.5 links). Zur Modellierung beider Kurven wurde anhand dieser die Funktion $f(x) = -0{,}14\sin(6{,}5x) + x$ abgeschätzt (vgl. Abbildung 3.5 rechts). Die Modellkalibrierung erfolgte entsprechend, indem anhand von $f(x)$ der vertikale Abstand jedes Vorhersagewertes zur Ursprungsgeraden g mit $g(x) = x$ berechnet wurde. Da der Vorhersagewert folglich um diesen Abstand von der Kalibrierungskurve eines perfekt kalibrierten Modell abweicht, wurde er durch Addition bzw. Subtraktion dieses Abstandes korrigiert. Zur Validierung dieser Kalibrierungstechnik wurde das beschriebene Verfahren mit identischer Funktion $f(x)$ auch auf den berechneten Vorhersagewerte basierend auf den Datensätzen der Testdaten angewandt.

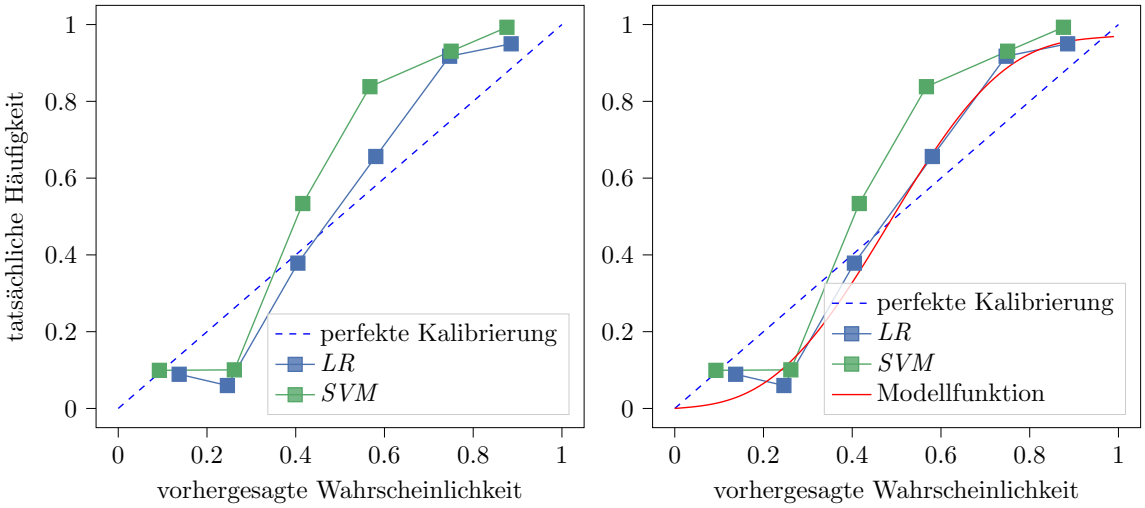

Abb. 3.5: **Kalibrierungskurven des multimodalen Classifiers für die mittels *LR* und *SVM* berechneten Vorhersagewerte für IPS.** Die Kurven zeigen, wie sich die vorhergesagte Wahrscheinlichkeit für IPS von der tatsächlichen Häufigkeit dieser Erkrankung innerhalb sechs unterschiedlichen Subklassen unterscheidet (links). Die Subklassen wurden gebildet, indem Probanden mit ähnlicher vorhergesagter Wahrscheinlichkeit für IPS zusammengefasst wurden. Beide Kalibrierungskurven lassen sich annähernd durch die Modellfunktion $f(x) = -0{,}14\sin(6{,}5x) + x$ modellieren (rechts).

3.3 Externe Validierung der Modelle

Für die externe Validierung der Modelle ist es erforderlich, dass die Testdaten von einer anderen Kohorte stammen als die Trainingsdaten. Während die interne Validierung zur Optimierung und Auswahl des Modells beiträgt, wird mit der externe Validierung überprüft, inwiefern es unter Verwendung dieses Modells möglich ist, auf Datensätze von außerhalb der eigenen Kohorte zu generalisieren und zuverlässige Vorhersagen zu berechnen (vgl. Abschnitt 5.5.2).

Der T1-, der rs-fMRT- sowie der multimodale Classifier wurden im Anschluss an das Training mit den Aufnahmen der HHU-Kohorte nach Matching ($n = 135$) an den Aufnahmen der externen Kohorten NEUROCON ($n = 43$) und Tao Wu ($n = 40$) getestet. Dies umfasste zunächst die Klassifizierung der T1-gewichteten sowie der rs-fMRT-Aufnahmen durch die mit den Datensätzen der HHU-Kohorte trainierten T1- und rs-fMRT-Classifier. Da eine strukturelle T1-gewichtete Aufnahme aus der NEUROCON-Kohorte nicht unter Verwendung von

CAT12 verarbeitet werden konnte, wurde der entsprechende Proband ausgeschlossen, sodass sich diese Kohorte auf $n = 42$ verkleinerte. Anschließend wurden die somit berechneten Klassifizierungsergebnisse in Kombination mit Angaben zum Alter und Geschlecht der Probanden verwendet, um mit Hilfe des mit den Datensätzen der HHU-Kohorte trainierten multimodalen Classifiers eine finale Gruppenvorhersage der Probanden zu berechnen. Die Vorverarbeitung der Aufnahmen bzw. der Klassifizierungsergebnisse sowie das Training der Modelle erfolgten analog zur internen Validierung der entsprechenden Classifier (vgl. Abschnitt 3.2.1, 3.2.2 und 3.2.4). Da bei jedem Trainingsdurchlauf die initialen Gewichte des Modells zufällig festgelegt werden, wurden alle Trainingsdurchläufe pro Classifier insgesamt zehnmal durchgeführt.

3.4 Untersuchung der Generalisierbarkeit auf multizentrische Kohorten

In den vorherigen Abschnitten wurden die Modelle ausschließlich mit Aufnahmen einer einzelnen Kohorte trainiert. Dies birgt das Risiko von *overfitting*, da die entsprechenden Modelle möglicherweise nur für diese spezifische Kohorte geeignet sind und fehlerhafte Vorhersagen ausgeben, wenn sie auf andere Populationen angewendet werden. Um die Generalisierbarkeit weiter zu verbessern, wurden die Modelle in den nachfolgenden Abschnitten mit den Datensätzen aus multizentrischen Kohorten trainiert. Da Trainings- und Testdaten nie aus derselben Kohorte stammten, wurde dabei ebenfalls die externe Validität der Modelle überprüft. Diese Form der Validierung hat darüber hinaus den Vorteil, dass eine große Anzahl an Datensätzen für das Training zur Verfügung steht und deren Heterogenität zunimmt. Dadurch verringert sich wiederum der störende Einfluss von *site effects* bzw. *scanner effects* (vgl. Abschnitt 5.5.2).

3.4.1 T1-Classifier bei idealer Alters- und Geschlechterverteilung

Der T1-Classifier wurde mit einer selbst zusammengestellten Kohorte ($n = 380$) bestehend aus ausgewählten Aufnahmen von Probanden der PPMI-, der ADNI- sowie der Rockland-Studie trainiert und anschließend an Aufnahmen aus der HHU-Kohorte ohne Matching ($n = 156$) getestet. Die Trainingskohorte umfasste 190 strukturelle T1-gewichtete Aufnahmen von Patienten mit IPS aus der PPMI-Studie und die gleiche Anzahl an Aufnahmen von GK aus der PPMI-, der ADNI- sowie der Rockland-Studie. Die Probanden wurden dabei so ausgewählt, dass eine identische Alters- und Geschlechterverteilung in beiden Gruppen erreicht wurde (vgl. Tabelle 3.9). Somit wurden diese Merkmale als mögliche Confounder weitgehend ausgeschlossen. Zwei von ursprünglich 192 verfügbaren Aufnahmen von Patienten mit IPS aus der PPMI-Kohorte konnten nicht gematcht werden und wurden daher verworfen. Die Vorverarbeitung der Aufnahmen sowie das Training des Modells erfolgten analog zur internen Validierung des T1-Classifiers (vgl. Abschnitt 3.2.1). Da bei jedem Trainingsdurchlauf die initialen Gewichte des Modells zufällig festgelegt werden, wurden alle Trainingsdurchläufe insgesamt zehnmal durchgeführt.

	Trainingsdaten (PPMI, ADNI, Rockland)			Testdaten (HHU)		
	n	davon Männer	Alter (Jahre)	n	davon Männer	Alter (Jahre)
IPS	190	120 (63 %)	$62{,}6 \pm 9{,}3$	78	50 (64 %)	$61{,}4 \pm 9{,}7$
GK	190	120 (63 %)	$62{,}6 \pm 9{,}3$	78	42 (54 %)	$55{,}4 \pm 10{,}6$
gesamt	380	240 (63 %)	$62{,}6 \pm 9{,}3$	156	92 (59 %)	$58{,}4 \pm 10{,}6$
p-Wert		1	1		0,25	$< 0{,}001$

Tabelle 3.9: **Trainings- und Testdaten der selbst zusammengestellten Kohorte** Die Trainingsdaten bestehend aus Aufnahmen von Probanden der PPMI-, der ADNI- sowie der Rockland-Studie wurden so zusammengesetzt, dass die Alters- und Geschlechterverteilung in beiden Gruppen identisch ist. Die p-Werte beziehen sich auf einen Zweistichproben-t-Test, der einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe untersucht bzw. auf einen χ^2-Unabhängigkeitstest für das Geschlecht in beiden Gruppen.

3.4.2 rs-fMRT-Classifier mit Leave-One-Site-Out-Ansatz

Bei diesem Ansatz wurden rs-fMRT-Aufnahmen aus der HHU-Kohorte nach Matching ($n = 135$) sowie der Aachen- ($n = 46$), der Köln- ($n = 26$), der NEUROCON- ($n = 43$) und der Tao Wu-Kohorte ($n = 40$) verwendet. Der rs-fMRT-Classifier wurde mit den Aufnahmen aus jeweils vier Kohorten trainiert und an den Aufnahmen der verbleibenden Kohorte getestet (‘*leave-one-site-out*’). Auf Grund einer unterschiedlichen Anzahl akquirierter *EPI-volumes* (vgl. Tabelle 3.5) wurde aus den Zeitserien der Probanden verschiedener Kohorten eine unterschiedliche Anzahl an *snippets*, und somit auch Datensätze, gebildet (vgl. Tabelle 3.10). Alle generierten *snippets* deckten jeweils 10 *time points* der akquirierter *EPI-volumes* ab. In Abhängigkeit von der Gesamtzahl an *time points* pro Aufnahme ergaben sich aus jeder Aufnahme 30 (HHU-Kohorte), 27 (Aachen-Kohorte), 18 (Köln-Kohorte), 13 (NEUROCON-Kohorte) bzw. 23 (Tao Wu-Kohorte) *snippets*. Die Vorverarbeitung der Aufnahmen sowie das Training des Modells verliefen analog zur internen Validierung des rs-fMRT-Classifiers (vgl. Abschnitt 3.2.2). Da bei jedem Trainingsdurchlauf die initialen Gewichte des Modells zufällig festgelegt werden, wurden alle Trainingsdurchläufe insgesamt zehnmal durchgeführt.

3.5 Modellbewertung

Zur Evaluation der Klassifizierungsleistung der Modelle wurden für jeden Ansatz die durchschnittliche *ROC AUC*, die *balanced accuracy*, die Sensitivität sowie die Spezifität (vgl. Abschnitt 2.1.5) aus insgesamt zehn Komplettdurchläufen inklusive deren 95 %-Konfidenzintervall unter der Annahme einer Normalverteilung berechnet. Diese Metriken wurden auf Probandenniveau bestimmt, sodass bei dem rs-fMRT-Classifier berücksichtigt werden musste, dass ein Proband als Patient mit IPS klassifiziert wurde, wenn IPS bei mindestens 50 % seiner *snippets* vorhergesagt wurde. Zu Beginn jedes Komplettdurchlaufs wurden die Modelle mit zufälligen Gewichten (*weights* und *biases*) initialisiert, zudem erfolgt bei der Verwendung der Kreuzvalidierung eine zufällige Partitionierung des Originaldatensatzes in Trainings- und Testdaten. Das zu jeder Metrik angegebene 95 %-Konfidenzintervall gibt den Bereich um einen Mittelwert der jeweiligen Metrik an, in dem der gesuchte Parameter mit einer Wahrscheinlichkeit von 95 % liegt.

3.6 Implementation und verwendete Software

Zur Implementation der beschriebenen *deep learning*-Verfahren wurde die Bibliothek Keras 2.3.0 (Chollet, 2015) mit TensorFlow 2.2.0 (Abadi et al., 2015) als *back-end* ausgewählt und innerhalb von Python 3.8.5 (Van Rossum und Drake, 2009) ausgeführt. Keras ermöglicht die

	Trainingsdaten			Testdaten		
	n	davon Männer	Alter (Jahre)	n	davon Männer	Alter (Jahre)
HHU						
IPS	1639	1086 (66 %)	$65{,}8 \pm 8{,}1$	1890	1050 (56 %)	$59{,}1 \pm 9{,}5$
GK	1550	940 (61 %)	$64{,}1 \pm 6{,}7$	2160	1230 (57 %)	$56{,}4 \pm 10{,}3$
gesamt	3189	2026 (64 %)	$65{,}0 \pm 7{,}5$	4050	2280 (56 %)	$57{,}7 \pm 10{,}0$
p-Wert		0,001	$< 0{,}001$			$< 0{,}001$
Aachen						
IPS	2935	1758 (60 %)	$61{,}6 \pm 9{,}5$	594	378 (64 %)	$65{,}3 \pm 8{,}5$
GK	3062	1792 (59 %)	$58{,}9 \pm 10{,}3$	648	378 (58 %)	$63{,}2 \pm 4{,}9$
gesamt	5997	3550 (59 %)	$60{,}2 \pm 10{,}0$	1242	756 (61 %)	$64{,}2 \pm 7{,}0$
p-Wert		0,29	$< 0{,}001$			$< 0{,}001$
Köln						
IPS	3295	1902 (58 %)	$62{,}1 \pm 9{,}6$	234	234 (100 %)	$63{,}5 \pm 7{,}6$
GK	3476	1936 (56 %)	$59{,}5 \pm 9{,}9$	234	234 (100 %)	$62{,}2 \pm 5{,}8$
gesamt	6771	3838 (57 %)	$60{,}8 \pm 9{,}8$	468	468 (100 %)	$62{,}9 \pm 6{,}8$
p-Wert		0,10	$< 0{,}001$			0,04
NEUROCON						
IPS	3178	1915 (60 %)	$61{,}5 \pm 9{,}1$	351	221 (63 %)	$68{,}7 \pm 10{,}4$
GK	3502	2118 (60 %)	$59{,}2 \pm 9{,}4$	208	52 (25 %)	$67{,}6 \pm 11{,}5$
gesamt	6680	4033 (60 %)	$60{,}3 \pm 9{,}3$	559	273 (49 %)	$68{,}3 \pm 10{,}8$
p-Wert		0,87	$< 0{,}001$			0,26
Tao Wu						
IPS	3069	1883 (61 %)	$61{,}8 \pm 9{,}9$	460	253 (55 %)	$65{,}2 \pm 4{,}3$
GK	3250	1894 (58 %)	$58{,}9 \pm 10{,}0$	460	276 (60 %)	$65{,}4 \pm 5{,}4$
gesamt	6319	3777 (60 %)	$60{,}3 \pm 10{,}0$	920	529 (58 %)	$65{,}4 \pm 4{,}9$
p-Wert		0,01	$< 0{,}001$			0,17

Tabelle 3.10: **Trainings- und Testdaten im *leave-one-site-out*-Ansatz.** Die Spalten mit der Überschrift 'Testdaten' enthalten die Zusammensetzung der jeweiligen getesteten Kohorte (fett gedruckt). Die Spalten mit der Überschrift 'Trainingsdaten' umfassen die Zusammensetzung der übrigen vier Kohorten. Da in Abhängigkeit von der Kohorte unterschiedlich viele *snippets* aus der Aufnahme eines einzelnen Probanden gebildet wurden, sind alle erhobenen Angaben auf *snippets*-Niveau anstelle von Probandenniveau. Die p-Werte beziehen sich auf einen Zweistichproben-t-Test, der einen signifikanten Altersunterschied zwischen IPS- und GK-Gruppe untersucht bzw. auf einen χ^2-Unabhängigkeitstest für das Geschlecht in beiden Gruppen.

Erstellung, das Training sowie die Validierung verschiedener *deep learning*-Modelle, zudem können kompatible Grafikkarten für die während dem Modelltraining anfallende Berechnungen genutzt werden. Es besteht dabei im Vergleich zum Training auf einem Hauptprozessor ein deutlicher Geschwindigkeitsvorteil, der insbesondere bei der Verarbeitung umfangreicher und komplexer Datenmengen ins Gewicht fällt (Lawrence et al., 2017).

Innerhalb von Python wurde die Bibliothek SciPy 1.4.1 (Virtanen et al., 2020) zur Berechnung der Konfidenzintervalle und der p-Werte der Zweistichproben-t-Tests sowie der χ^2-Unabhängigkeitstests verwendet. Bei der Modellkalibrierung, der Umsetzung der Kreuzvalidierung und der Anwendung von *LR* sowie *SVM* wurde auf scikit-learn 0.24.1 (Pedregosa et al., 2011) zurückgegriffen. Grafiken wurden u. a. mit seaborn 0.11.1 (Waskom, 2021) bzw. Matplotlib 3.3.4 (Hunter, 2007) erstellt und mit tikzplotlib 0.9.8 (`pypi.org/project/tikzplotlib/`) als Vektorgrafik exportiert. Die Erstellung, das Bearbeiten, das Exportieren und das Laden aller MRT-Aufnahmen wurde innerhalb von Python mit SimpleITK 2.0.2

(Lowekamp et al., 2013), Nipype 1.6.0 (Gorgolewski et al., 2016) und NiBabel 3.2.1 (Brett et al., 2020) durchgeführt. NumPy 1.19.0 (Harris et al., 2020) bzw. pandas 1.2.3 (Reback et al., 2021) wurden für die Verarbeitung von Arrays bzw. Tabellen genutzt.

Außerhalb von Python wurden die MRT-Aufnahmen aus den externen Kohorten mit dcm2niix 1.0.20190902 (X. Li et al., 2016) vom DICOM- ins NIfTI-Format konvertiert. Die allgemeine Vorverarbeitung der Aufnahmen erfolgte mit FSL 6.0.4 (Woolrich et al., 2009; S. M. Smith et al., 2004; Jenkinson et al., 2012) sowie CAT12.7 (Gaser und Dahnke, 2016). Zur Darstellung von Aufnahmen, *IC-maps* und *heatmaps* wurde FSLeyes 0.34.2 (McCarthy, 2020) genutzt. Die Abbildungen 1.1, 2.1, 2.9, 3.1, 3.2 und 3.4 wurden unter Verwendung der Vektorgrafiksoftware Inkscape 1.1.1 (`inkscape.org`) erstellt.

Für die Durchführung der meisten Berechnungen wurden ein AMD® Ryzen 7 3700X mit 64 Gigabyte Arbeitsspeicher und eine NVIDIA® GeForce RTX 2080 Ti mit 11 Gigabyte Grafikspeicher genutzt. Eine Ausnahme davon bildete die Vorverarbeitung der strukturellen T1-gewichteten MRT-Aufnahmen mit CAT12.7, die auf dem High-Performance-Computing-Cluster ‘HILBERT’ der HHU durchgeführt wurde.

4 Ergebnisse

4.1 Uni- und multimodale Modelle auf interner Kohorte

In einem ersten Schritt wurde die interne Validität der vier beschriebenen Classifier untersucht. Dazu wurden diese ausschließlich mit den Datensätzen der Probanden aus der HHU-Kohorte nach Matching trainiert und anschließend an den übrigen Datensätzen aus dieser Kohorte validiert.

4.1.1 T1-Classifier

Der T1-Classifier war nach entsprechendem Training in der Lage, bei den Probanden aus den Testdaten ($n = 27$) der HHU-Kohorte nach Matching zwischen Patienten mit IPS und GK zu differenzieren. Die *ROC AUC* lag bei 0,81 [0,77; 0,84], die *balanced accuracy* bei 71,2 % [68,5 %; 73,9 %], die Sensitivität bei 73,1 % [67,7 %; 78,4 %] und die Spezifität bei 69,3 % [63,9 %; 74,7 %] (vgl. Tabelle 4.1). Die berechneten *class activation maps* korrekt vorhergesagter Patienten mit IPS sind in Abbildung 4.1 in Form von *heatmaps* dargestellt und werden in Unterkapitel 5.4 diskutiert.

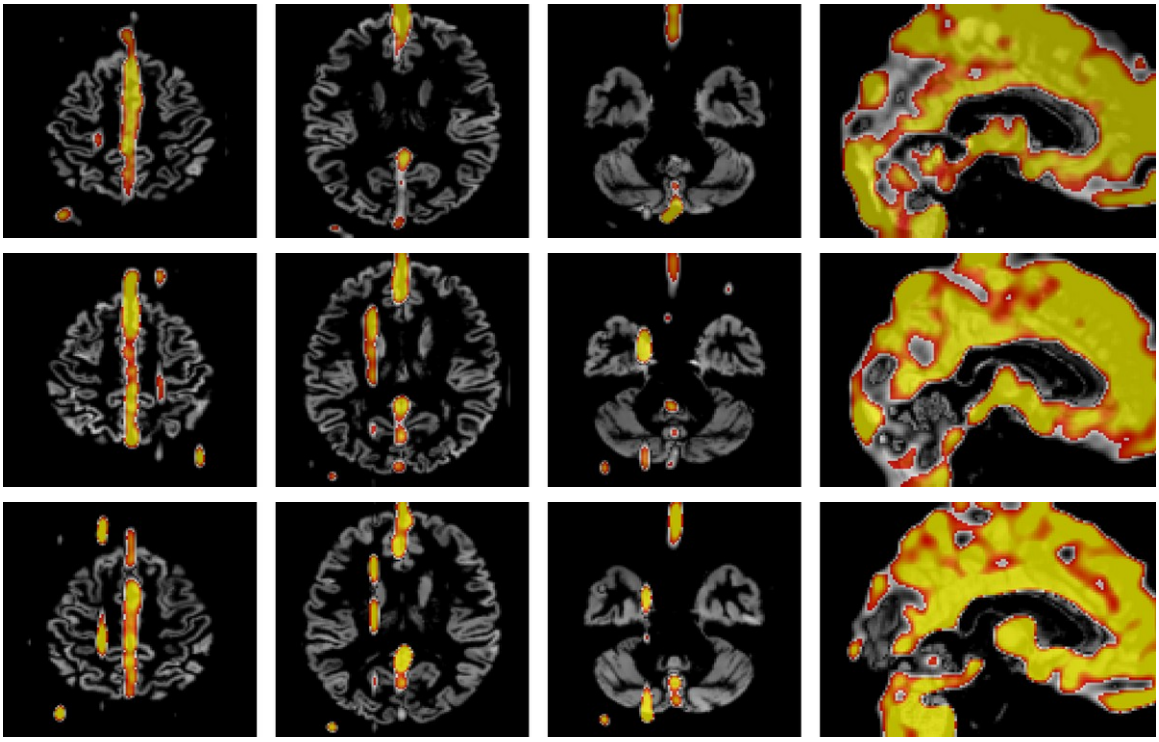

Abb. 4.1: **Beispielhafte *heatmaps* von drei Patienten mit IPS aus den Testdaten der HHU-Kohorte nach Matching.** Pro Patient sind drei axiale und eine sagittale Rekonstruktion abgebildet. Die farbigen Anteile repräsentieren die Regionen der vorverarbeiteten Aufnahme, die zur korrekten Vorhersage geführt haben. Für die Berechnung dieser Vorhersage wurden gelb markierte Bereiche vom Modell stärker gewichtet als solche, die rot markiert sind.

Diese Ergebnisse konnten im Rahmen der 5-fachen Kreuzvalidierung an der gesamten Ko-

horte nach Matching ($n = 135$) nur teilweise generalisiert werden: Die *ROC AUC* war mit 0,69 [0,67; 0,71] signifikant niedriger. Eine ähnliche Einbuße zeigte sich bei der *balanced accuracy* mit 62,1 % [60,2 %; 63,9 %] und der Sensitivität mit 54,3 % [51,1 %; 57,4 %], nicht aber bei der Spezifität, die mit 69,9 % [64,7 %; 75,0 %] fast unverändert blieb (vgl. Tabelle 4.2 und 4.3).

4.1.2 rs-fMRT-Classifier

Mit einer *ROC AUC* von 0,94 [0,93; 0,95] sowie einer *balanced accuracy* von 83,7 % [81,1 %; 86,2 %] zeigte der rs-fMRT-Classifier eine hohe Genauigkeit bei der Zuordnung von Aufnahmen zu Patienten mit IPS bzw. GK bei den Probanden aus den Testdaten ($n = 27$) der HHU-Kohorte nach Matching. Die Sensitivität lag bei 72,3 % [67,0 %; 77,6 %] und die Spezifität bei 95,0 % [92,5 %; 97,5 %] (vgl. Tabelle 4.1).

Mittels 5-facher Kreuzvalidierung an der gesamten Kohorte nach Matching ($n = 135$) konnten diese Ergebnisse weitgehend reproduziert werden. Die *ROC AUC* betrug 0,90 [0,89; 0,91], die *balanced accuracy* 82,7 % [81,5 %; 83,9 %], die Sensitivität 76,2 % [74,3 %; 78,0 %] und die Spezifität 89,2 % [87,3 %; 91,0 %] (vgl. Tabelle 4.2 und 4.3). Des Weiteren zeigte sich nach dualer Regression unter Verwendung einer aus weichgezeichnetem Rauschen bestehenden *pseudo group IC-map* als Negativkontrolle eine signifikante Verschlechterung der *ROC AUC* auf 0,73 [0,71; 0,75]. Die *balanced accuracy* betrug dabei 64,2 % [61,7 %; 66,7 %], die Sensitivität 41,9 % [40,0 %; 43,8 %] und die Spezifität 86,4 % [83,3 %; 89,5 %].

4.1.3 DTI-Classifier

Für den DTI-Classifier standen weniger Trainings- ($n = 84$) und Testdaten ($n = 22$) als für die anderen Classifiern zur Verfügung, da von einzelnen Probanden keine diffusionsgewichteten Aufnahmen akquiriert wurden (vgl. Abschnitt 3.1.1). Mit einer *ROC AUC* von 0,75 [0,73; 0,77] und einer *balanced accuracy* von 66,8 % [63,3 %; 70,3 %] konnte dieser bei der HHU-Kohorte nach Matching jedoch ebenfalls anhand entsprechender Aufnahmen überwiegend korrekt zwischen Patienten mit IPS und GK differenzieren. Die Sensitivität schwankte erheblich und lag durchschnittlich bei 76,9 % [62,0 %; 91,8 %]. Ähnlich variierte die Spezifität mit 56,7 % [41,0 %; 72,3 %] (vgl. Tabelle 4.1).

Im Rahmen der 5-fachen Kreuzvalidierung war die Kohorte nach Matching ($n = 106$) erneut auf Grund fehlender Aufnahmen anders zusammengesetzt (vgl. Abschnitt 3.1.1). Es ergaben sich eine *ROC AUC* von 0,70 [0,68; 0,73], eine *balanced accuracy* von 64,0 % [62,2 %; 65,9 %], eine Sensitivität von 66,8 % [59,2 %; 74,5 %] und eine Spezifität von 61,2 % [53,3 %; 69,2 %] (vgl. Tabelle 4.2 und 4.3).

4.1.4 Multimodaler Classifier

Zur Evaluation des multimodalen Classifiers wurden die mittels *LR* und *SVM* berechneten Vorhersagen miteinander verglichen. Nach Anwendung von *LR* erreichte die *ROC AUC* für die Gruppenvorhersage der Probanden aus den Testdaten ($n = 27$) der HHU-Kohorte nach Matching einen Wert von 0,96 [0,95; 0,97]. Die *balanced accuracy* lag bei 89,6 % [87,4 %; 91,8 %], die Sensitivität bei 89,2 % [83,9 %; 94,5 %] und die Spezifität bei 90,0 % [87,4 %; 92,6 %]. Entsprechend betrugen die *ROC AUC* nach Anwendung von *SVM* 0,96 [0,95; 0,96], die *balanced accuracy* 88,2 % [85,7 %; 90,7 %], die Sensitivität 89,2 % [83,9 %; 94,5 %] und die Spezifität 87,1 % [83,9 %; 90,4 %] (vgl. Tabelle 4.1). Im Hinblick auf die *balanced accuracy* sowie die Sensitivität war folglich die Leistung des multimodalen Classifiers signifikant besser als die des rs-fMRT-Classifiers.

	Männer						Frauen					
	Alter (Jahre)	T1 (%)	rs-fMRT (%)	DTI (%)	multimodal *LR* (%)	multimodal *SVM* (%)	Alter (Jahre)	T1 (%)	rs-fMRT (%)	DTI (%)	multimodal *LR* (%)	multimodal *SVM* (%)
IPS	39	46	60	48	57	70	55	75	77	69	88	98
	50	25	26	76	31	21	57	65	61	54	74	90
	61	97	48	46	61	69	59	94	16	73	57	67
	64	30	98	87	85	95	63	97	93	34	91	99
	66	28	73	69	65	71	68	92	53	97	85	95
	80	94	100	90	94	99	75	91	94	77	94	99
							78	72	72	77	85	95
GK	41	8	9	42	11	4	27	2	5	fehlt	17	13
	44	75	1	74	29	20	42	8	26	59	32	32
	51	11	56	93	56	61	56	91	31	46	59	73
	55	0	10	fehlt	11	3	62	33	1	21	13	5
	58	80	23	16	26	16	63	22	17	8	16	7
	69	34	15	fehlt	18	6	76	69	3	47	27	15
	75	5	11	fehlt	11	2	81	39	9	fehlt	23	10

Tabelle 4.1: **Vorhergesagte Wahrscheinlichkeit für IPS in Prozent bei Probanden aus den Testdaten (*80:20 split*) der HHU-Kohorte nach Matching getrennt für jede Bildgebungsmodalität und multimodal.** Das Training erfolgte mit den Trainingsdaten der HHU-Kohorte nach Matching. Angegeben ist jeweils das arithmetische Mittel aus zehn Iterationen. Der Cut-off-Wert zur Zuordnung eines Klassifizierungsergebnisses zu einem Patienten mit IPS oder einem GK liegt bei 50 %, korrekte bzw. inkorrekte Vorhersagen sind grün bzw. rot unterlegt.

Im Rahmen der 5-fachen Kreuzvalidierung an der gesamten Kohorte nach Matching ($n = 135$) zeigte sich erwartungsgemäß eine geringe Abnahme der Klassifizierungsgenauigkeit. Nach Anwendung von *LR* für die abschließende Vorhersage der Gruppenzugehörigkeit belief sich die *ROC AUC* auf 0,91 [0,90; 0,92], die *balanced accuracy* auf 84,2 % [82,7 %; 85,6 %], die Sensitivität auf 80,0 % [78,1 %; 81,9 %] und die Spezifität auf 88,5 % [87,2 %; 89,7 %]. Nach Klassifizierung mittels *SVM* lag die *ROC AUC* bei 0,91 [0,91; 0,92], die *balanced accuracy* bei 85,1 % [83,8 %; 86,3 %], die Sensitivität bei 82,7 % [80,7 %; 84,7 %] und die Spezifität bei 87,5 % [85,7 %; 89,3 %] (vgl. Tabelle 4.2 und 4.3). Der multimodale Classifier zeigte somit erneut die beste Klassifizierungsleistung, jedoch war diese im Rahmen der 5-fachen Kreuzvalidierung nicht signifikant besser als die des rs-fMRT-Classifiers. Eine Übersicht über die Vorhersagegenauigkeit der verwendeten Classifier ist in Abbildung 4.2 sowie in Tabelle 4.4 dargestellt.

Zusätzlich wurde analysiert, wie sich die Vorhersagegenauigkeit ändert, wenn der multimodale Classifier ausschließlich mit den Klassifizierungsergebnissen aller unimodalen Classifier (Variante 1) oder nur mit den Klassifizierungsergebnissen des rs-fMRT-Classifiers in Verbindung mit Angaben zum Alter und Geschlecht der Probanden (Variante 2) trainiert wurde. Im Vergleich dazu berücksichtigte der multimodale Classifier in seiner unveränderten Form sowohl die Klassifizierungsergebnissen aller unimodalen Classifier als auch das Alter sowie das Geschlecht der Probanden (Referenz). Es zeigte sich, dass der multimodale Classifier unter Verwendung möglichst vieler Eingabedaten den beiden anderen Varianten tendenziell überlegen war (vgl. Tabelle 4.5).

Die anschließende Modellkalibrierung erfolgte gemäß der in Abschnitt 3.2.4 beschriebenen Methode. Sie wurde anhand der Datensätze zufällig ausgewählter Probanden ($n = 95$) aus der HHU-Kohorte nach Matching entwickelt und anschließend auf diese Datensätze angewandt. Der *Brier score* für die Vorhersagen des multimodalen Classifiers nach Anwendung von *LR* verringerte sich dadurch von 0,152 [0,146; 0,158] um 0,008 [0,005; 0,011] auf 0,144 [0,136; 0,153] Punkte. Hingegen stieg der *Brier score* für die Vorhersagen des multimoda-

	Männer						Frauen					
	Alter (Jahre)	T1 (%)	rs-fMRT (%)	DTI (%)	multimodal *LR* (%)	multimodal *SVM* (%)	Alter (Jahre)	T1 (%)	rs-fMRT (%)	DTI (%)	multimodal *LR* (%)	multimodal *SVM* (%)
IPS	39	30	64	31	56	64	41	2	37	fehlt	29	28
	44	11	25	80	27	18	44	15	79	29	62	85
	45	17	88	47	77	93	47	9	24	26	23	15
	45	43	74	45	66	76	48	41	58	26	58	68
	50	35	29	57	30	22	49	10	78	39	76	88
	51	53	75	94	84	97	50	13	66	fehlt	57	75
	51	28	63	71	61	72	51	48	84	45	80	93
	51	2	97	55	77	90	53	43	47	51	56	65
	52	15	46	90	46	49	53	76	62	79	79	92
	52	10	82	26	58	66	55	47	82	63	82	94
	53	1	79	42	56	70	57	36	58	49	61	73
	54	61	14	90	32	23	58	7	27	47	25	12
	55	93	83	fehlt	87	98	59	50	56	65	68	83
	55	1	94	42	72	89	59	94	20	53	41	36
	57	72	93	39	84	97	60	33	54	57	58	66
	57	44	58	fehlt	57	66	62	99	83	90	93	98
	57	52	71	78	73	85	62	25	78	43	69	85
	59	19	83	fehlt	68	79	63	61	93	20	83	97
	59	58	65	44	60	69	65	94	63	70	82	89
	60	58	100	87	91	98	68	87	63	93	88	97
	61	96	57	49	62	62	68	100	43	61	66	70
	61	94	98	90	93	98	68	75	45	86	67	67
	62	90	78	90	85	95	70	78	93	50	90	97
	62	88	48	82	63	68	71	75	3	fehlt	21	7
	62	44	99	67	83	95	72	93	54	39	64	70
	63	96	78	76	86	98	72	87	90	59	88	97
	64	14	99	84	85	98	75	79	90	80	90	98
	64	52	94	81	84	95	78	88	76	68	85	97
	65	66	97	86	92	99						
	66	35	71	64	61	73						
	67	98	88	85	93	99						
	73	96	96	83	92	99						
	76	95	98	86	93	99						
	76	71	77	70	77	89						
	80	89	100	77	88	97						

Tabelle 4.2: **Vorhergesagte Wahrscheinlichkeit für IPS in Prozent bei Patienten mit IPS aus der HHU-Kohorte nach Matching getrennt für jede Bildgebungsmodalität und multimodal.** Angegeben ist jeweils das arithmetische Mittel aus zehn Iterationen von Klassifizierungen mittels 5-facher Kreuzvalidierung. Der Cut-off-Wert zur Zuordnung eines Klassifizierungsergebnisses zu einem Patienten mit IPS oder einem GK liegt bei 50 %, korrekte bzw. inkorrekte Vorhersagen sind grün bzw. rot unterlegt, uneindeutige Vorhersagen (exakt 50 %) sind gelb markiert.

	Männer						Frauen					
	Alter (Jahre)	T1 (%)	rs-fMRT (%)	DTI (%)	multimodal *LR* (%)	multimodal *SVM* (%)	Alter (Jahre)	T1 (%)	rs-fMRT (%)	DTI (%)	multimodal *LR* (%)	multimodal *SVM* (%)
GK	37	38	7	fehlt	17	7	27	0	7	fehlt	12	6
	41	6	27	17	16	7	42	13	19	47	25	22
	41	86	31	41	45	50	43	5	20	64	31	24
	41	51	15	57	23	13	44	2	36	16	28	24
	41	5	15	52	17	8	50	0	90	fehlt	83	95
	44	82	4	76	33	18	51	12	12	17	15	6
	44	4	79	14	61	81	51	2	37	fehlt	34	25
	45	7	2	25	8	2	51	4	1	25	10	5
	46	81	67	50	76	92	54	18	16	31	18	5
	47	79	9	43	25	12	55	23	7	23	14	6
	48	0	30	fehlt	23	16	56	1	2	fehlt	14	3
	49	19	9	76	19	7	56	23	7	44	16	6
	49	92	12	54	29	17	56	65	30	49	46	42
	51	38	50	93	62	73	57	2	29	fehlt	27	20
	51	21	6	33	11	3	58	1	3	fehlt	12	2
	52	79	6	38	19	8	59	2	26	64	27	15
	52	23	11	68	19	9	59	12	78	fehlt	71	88
	52	38	29	42	28	17	60	18	2	37	13	4
	53	60	5	fehlt	18	5	62	52	39	49	47	50
	53	8	47	69	41	42	62	66	1	22	19	7
	55	1	17	fehlt	15	7	62	41	8	fehlt	22	7
	55	1	30	fehlt	21	8	62	41	2	25	16	4
	55	36	35	37	28	18	62	29	30	fehlt	35	33
	56	17	4	18	8	2	63	1	6	fehlt	12	2
	58	75	24	25	32	21	63	33	12	15	18	6
	59	1	6	fehlt	9	1	65	76	31	11	32	21
	60	44	37	41	36	27	65	21	26	fehlt	30	26
	60	79	22	50	31	19	68	3	5	33	11	2
	60	94	18	22	32	20	70	65	54	79	75	88
	60	39	4	fehlt	13	4	76	74	6	48	24	7
	61	8	5	42	10	3	81	64	20	fehlt	32	14
	62	46	52	29	45	48						
	63	31	16	31	17	5						
	66	29	32	fehlt	27	16						
	69	28	22	55	22	11						
	69	60	19	fehlt	22	9						
	70	27	24	86	35	18						
	71	15	10	54	10	3						
	73	4	56	fehlt	32	30						
	75	6	22	fehlt	17	7						
	78	47	13	84	33	13						

Tabelle 4.3: **Vorhergesagte Wahrscheinlichkeit für IPS in Prozent bei GK aus der HHU-Kohorte nach Matching getrennt für jede Bildgebungsmodalität und multimodal.** Angegeben ist jeweils das arithmetische Mittel aus zehn Iterationen von Klassifizierungen mittels 5-facher Kreuzvalidierung. Der Cut-off-Wert zur Zuordnung eines Klassifizierungsergebnisses zu einem Patienten mit IPS oder einem GK liegt bei 50 %, korrekte bzw. inkorrekte Vorhersagen sind grün bzw. rot unterlegt, uneindeutige Vorhersagen (exakt 50 %) sind gelb markiert.

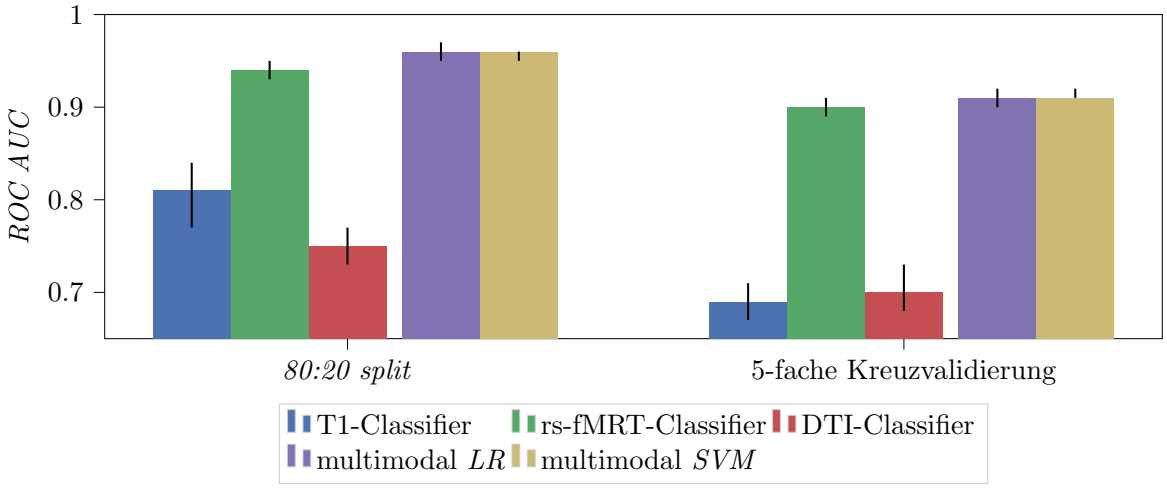

Abb. 4.2: **Vergleich aller Classifier bei Anwendung innerhalb der HHU-Kohorte nach Matching im Rahmen des *80:20 split* und der 5-fachen Kreuzvalidierung.** Angegeben sind jeweils die *ROC AUC* sowie das zugehörige 95 %-Konfidenzintervall.

Classifier	n	*ROC AUC*	*balanced accuracy* (%)	Sensitivität (%)	Spezifität (%)
80:20 split					
T1	27	0,81 [0,77; 0,84]	71,2 [68,5; 73,9]	73,1 [67,7; 78,4]	69,3 [63,9; 74,7]
rs-fMRT	27	0,94 [0,93; 0,95]	83,7 [81,1; 86,2]	72,3 [67,0; 77,6]	95,0 [92,5; 97,5]
DTI	22	0,75 [0,73; 0,77]	66,8 [63,3; 70,3]	76,9 [62,0; 91,8]	56,7 [41,0; 72,3]
multimodal *LR*	27	0,96 [0,95; 0,97]	89,6 [87,4; 91,8]	89,2 [83,9; 94,5]	90,0 [87,4; 92,6]
multimodal *SVM*	27	0,96 [0,95; 0,96]	88,2 [85,7; 90,7]	89,2 [83,9; 94,5]	87,1 [83,9; 90,4]
5-fache Kreuzvalidierung					
T1	135	0,69 [0,67; 0,71]	62,1 [60,2; 63,9]	54,3 [51,1; 57,4]	69,9 [64,7; 75,0]
rs-fMRT	135	0,90 [0,89; 0,91]	82,7 [81,5; 83,9]	76,2 [74,3; 78,0]	89,2 [87,3; 91,0]
DTI	106	0,70 [0,68; 0,73]	64,0 [62,2; 65,9]	66,8 [59,2; 74,5]	61,2 [53,3; 69,2]
multimodal *LR*	135	0,91 [0,90; 0,92]	84,2 [82,7; 85,6]	80,0 [78,1; 81,9]	88,5 [87,2; 89,7]
multimodal *SVM*	135	0,91 [0,91; 0,92]	85,1 [83,8; 86,3]	82,7 [80,7; 84,7]	87,5 [85,7; 89,3]

Tabelle 4.4: **Vergleich aller Classifier bei Anwendung innerhalb der HHU-Kohorte nach Matching im Rahmen des *80:20 split* und der 5-fachen Kreuzvalidierung.**

len Classifiers nach Anwendung von *SVM* von 0,141 [0,133; 0,149] um 0,003 [−0,001; 0,007] auf 0,144 [0,134; 0,155] Punkte an. Zur Testung der Methode zur Modellkalibrierung wurde sie ohne weitere Modifikationen auch auf die Datensätze der übrigen Probanden ($n = 40$) angewandt. Der *Brier score* für die Vorhersagen des multimodalen Classifiers nach Anwendung von *LR* verringerte sich damit von 0,103 [0,095; 0,111] um 0,022 [0,019; 0,026] auf 0,081 [0,070; 0,092] Punkte. Ebenfalls verringerte sich der *Brier score* für die Vorhersagen nach Anwendung von *SVM* von 0,086 [0,075; 0,097] um 0,009 [0,005; 0,014] auf 0,077 [0,063; 0,090] Punkte.

4.2 Externe Validierung der Modelle

In einem weiteren Schritt wurde die externe Validierung der Modelle durchgeführt. Um generell sicherzustellen, dass Modelle nicht für spezifische *features* in den Trainingsdaten optimiert

		ROC AUC	*balanced accuracy* (%)	Sensitivität (%)	Spezifität (%)
80:20 split					
Referenz	*LR*	0,957 [0,946; 0,967]	89,62 [87,44; 91,79]	89,23 [83,91; 94,55]	90,00 [87,36; 92,64]
	SVM	0,956 [0,947; 0,965]	88,19 [85,69; 90,68]	89,23 [83,91; 94,55]	87,14 [83,91; 90,37]
Variante 1	*LR*	0,963 [0,956; 0,970]	88,43 [86,06; 90,80]	86,15 [81,10; 91,21]	90,71 [87,27; 94,16]
	SVM	0,963 [0,955; 0,971]	86,90 [84,59; 89,20]	83,08 [78,02; 88,13]	90,71 [87,27; 94,16]
Variante 2	*LR*	0,926 [0,917; 0,935]	85,63 [83,19; 88,07]	77,69 [73,63; 81,75]	93,57 [89,80; 97,34]
	SVM	0,930 [0,918; 0,941]	85,66 [84,02; 87,30]	78,46 [74,98; 81,94]	92,86 [88,69; 97,03]
5-fache Kreuzvalidierung					
Referenz	*LR*	0,912 [0,903; 0,921]	84,19 [82,74; 85,64]	80,00 [78,13; 81,87]	88,47 [87,23; 89,72]
	SVM	0,915 [0,906; 0,924]	85,06 [83,76; 86,35]	82,70 [80,74; 84,66]	87,50 [85,75; 89,25]
Variante 1	*LR*	0,909 [0,900; 0,919]	83,81 [81,96; 85,66]	79,68 [76,86; 82,50]	87,78 [86,39; 89,17]
	SVM	0,910 [0,902; 0,918]	84,47 [82,79; 86,15]	81,43 [78,80; 84,05]	87,36 [85,85; 88,88]
Variante 2	*LR*	0,892 [0,883; 0,901]	83,33 [81,41; 85,25]	79,84 [77,83; 81,85]	86,94 [84,69; 89,20]
	SVM	0,902 [0,894; 0,910]	83,54 [81,44; 85,65]	80,95 [78,68; 83,22]	86,39 [83,29; 89,49]

Tabelle 4.5: **Evaluation des multimodalen Classifiers nach Training mit unterschiedlichen Varianten an Eingabedaten bei Anwendung innerhalb der HHU-Kohorte nach Matching im Rahmen des *80:20 split* und der 5-fachen Kreuzvalidierung.** In der ursprünglichen Methode (Referenz) erfolgten Training und Validierung unter Verwendung der Klassifizierungsergebnisse aller Bildgebungsmodalitäten in Kombination mit Angaben zum Alter und Geschlecht der Probanden. Im Gegensatz dazu wurden in Variante 1 die Angaben zum Alter und Geschlecht nicht berücksichtigt. In Variante 2 hingegen wurde der multimodale Classifier lediglich mit den Klassifizierungsergebnissen des rs-fMRT-Classifiers in Kombination mit Angaben zum Alter und Geschlecht der Probanden trainiert und anschließend validiert. In Abhängigkeit von der untersuchten Metrik sind die jeweiligen Werte des tendenziell überlegenen Vorgehens grün unterlegt.

sind, ist es erforderlich, dass Trainings- und Testdaten aus unterschiedlichen Kohorten stammen.

Zunächst wurden der T1-, der rs-fMRT- und der multimodale Classifier mit den entsprechenden Datensätzen der HHU-Kohorte nach Matching ($n = 135$) trainiert. Anschließend wurden die Classifier verwendet, um die Klassifizierungsergebnisse der Probanden aus der NEUROCON- ($n = 43$) und der Tao Wu-Kohorte ($n = 40$) anhand deren strukturellen T1-gewichteten sowie funktionellen Aufnahmen zu berechnen. Basierend auf diesen Klassifizierungsergebnissen sowie Angaben zum Alter und Geschlecht der Probanden wurden unter Verwendung des bereits trainierten multimodalen Classifiers schließlich die Gruppenvorhersagen für jeden Probanden berechnet. Es zeigte sich dabei, dass eine sichere Zuordnung der Probanden zu Patienten mit IPS bzw. GK mit keinem der verwendeten Classifier möglich war. In Tabelle 4.6 ist die Klassifizierungsleistung jedes Classifiers getrennt für beide Kohorten zusammengefasst.

4.3 Untersuchung der Generalisierbarkeit auf multizentrische Kohorten

Abschließend wurde untersucht, inwieweit mit Hilfe des T1- und des rs-fMRT-Classifiers zwischen Patienten mit IPS und GK differenziert werden kann, nachdem diese mit den Datensätzen von mehr als zwei Kohorten trainiert wurden.

Classifier	n	*ROC AUC*	*balanced accuracy* (%)	Sensitivität (%)	Spezifität (%)
NEUROCON					
T1	42	0,62 [0,60; 0,64]	55,7 [53,6; 57,8]	65,8 [59,4; 72,2]	45,6 [37,4; 53,8]
rs-fMRT	43	0,67 [0,64; 0,70]	62,6 [59,3; 65,8]	57,0 [52,9; 61,2]	68,1 [61,6; 74,6]
multimodal *LR*	42	0,64 [0,61; 0,66]	59,4 [54,2; 64,7]	58,8 [54,3; 63,3]	60,0 [49,2; 70,8]
multimodal *SVM*	42	0,64 [0,62; 0,67]	58,8 [53,2; 64,3]	56,9 [52,1; 61,7]	60,6 [49,9; 71,4]
Tao Wu					
T1	40	0,53 [0,50; 0,56]	53,0 [48,6; 57,4]	37,0 [25,9; 48,1]	69,0 [60,8; 77,2]
rs-fMRT	40	0,61 [0,58; 0,64]	55,3 [53,3; 57,2]	27,0 [22,2; 31,8]	83,5 [78,7; 88,3]
multimodal *LR*	40	0,59 [0,56; 0,63]	54,3 [51,7; 56,8]	23,5 [19,4; 27,6]	85,0 [78,0; 92,0]
multimodal *SVM*	40	0,59 [0,56; 0,63]	54,8 [52,2; 57,3]	23,5 [18,2; 28,8]	86,0 [80,0; 92,0]

Tabelle 4.6: **Klassifizierungsleistung der T1-, rs-fMRT- und multimodalen Classifier bei den externen Kohorten NEUROCON und Tao Wu.** Die Trainingsdaten stammen jeweils aus der HHU-Kohorte nach Matching. Metriken, deren 95 %-Konfidenzintervall einen Wert von 0,5 bzw. 50 % oder weniger einschließt, sind rot unterlegt dargestellt.

4.3.1 T1-Classifier bei idealer Alters- und Geschlechterverteilung

Nach dem Training des T1-Classifiers mit den Aufnahmen der zusammengesetzten Kohorte ($n = 390$) bestehend aus Probanden der PPMI-, der ADNI- und der Rockland-Studie wurde dieser an den Aufnahmen aus der HHU-Kohorte ohne Matching ($n = 156$) getestet. Es zeigte sich, dass dieser nicht für eine sichere Differenzierung zwischen Patienten mit IPS und GK geeignet war. Die *ROC AUC* lag bei 0,59 [0,57; 0,61], die *balanced accuracy* bei 56,7 % [55,2 %; 58,3 %], die Sensitivität bei 51,9 % [41,9 %; 62,0 %] und die Spezifität bei 61,5 % [51,8 %; 71,3 %].

4.3.2 rs-fMRT-Classifier mit Leave-One-Site-Out-Ansatz

Der rs-fMRT-Classifier wurde mit den *snippets* aus den Aufnahmen der Probanden aus vier Kohorten trainiert und an den *snippets* der Probanden aus der jeweils verbleibenden Kohorte getestet. Die berechneten Klassifizierungsmetriken zeigten, dass auch mit diesem Ansatz keine sichere Vorhersage der Gruppenzugehörigkeit eines Probanden möglich war (vgl. Tabelle 4.7).

Testkohorte	n	*ROC AUC*	*balanced accuracy* (%)	Sensitivität (%)	Spezifität (%)
HHU	135	0,71 [0,67; 0,74]	56,9 [54,2; 59,6]	18,6 [12,9; 24,2]	95,1 [94,1; 96,2]
Aachen	46	0,65 [0,63; 0,67]	58,0 [55,4; 60,6]	66,4 [62,9; 69,9]	49,6 [44,4; 54,7]
Köln	26	0,77 [0,73; 0,81]	68,1 [63,6; 72,6]	66,2 [59,2; 73,1]	70,0 [61,2; 78,8]
NEUROCON	43	0,66 [0,62; 0,69]	61,5 [57,5; 65,6]	73,7 [67,7; 79,8]	49,4 [38,1; 60,6]
Tao Wu	40	0,64 [0,60; 0,69]	61,7 [57,4; 66,1]	68,0 [59,9; 76,1]	55,5 [50,3; 60,7]

Tabelle 4.7: **Klassifizierungsleistung des rs-fMRT-Classifiers mit *leave-one-site-out*-Ansatz.** Die Trainingskohorten sind jeweils jene Kohorten, die nicht Testkohorte sind. Metriken, deren 95 %-Konfidenzintervall einen Wert von 0,5 bzw. 50 % oder weniger einschließt, sind rot unterlegt dargestellt.

5 Diskussion

5.1 Überblick

In der vorliegenden Studie konnte anhand der HHU-Kohorte gezeigt werden, dass eine Differenzierung zwischen Patienten mit IPS und GK anhand unterschiedlicher MRT-Bildgebungsmodalitäten nach spezifischer Vorverarbeitung mittels *deep learning*-Verfahren möglich ist. Dies wurde insbesondere bei der Analyse von rs-fMRT-Aufnahmen der Probanden erreicht. Durch den zusätzlichen Einbezug von strukturellen und diffusionsgewichteten Aufnahmen in Kombination mit Angaben zum Alter und Geschlecht der Probanden konnte die Vorhersagegenauigkeit weiter gesteigert werden. Es wurde dabei kein signifikanter Unterschied zwischen der Berechnung des finalen Klassifizierungsergebnisses mittels *LR* und *SVM* festgestellt.

Bei Betrachtung der Altersverteilung der Probanden aus der HHU-Kohorte fällt auf, dass vor allem Patienten mit IPS unter 60 Jahren fast unabhängig von ihrem Geschlecht oder dem verwendeten Classifier auffallend häufig als GK klassifiziert wurden (vgl. Tabelle 4.2). Dies wurde insbesondere bei den Vorhersagen des T1-Classifiers deutlich und legt den Schluss nahe, dass der Confounder Probandenalter möglicherweise nicht ausreichend kontrolliert wurde. Ramírez et al. (2019) stellten bei einem Vergleich der strukturellen Aufnahmen von de-novo IPS-Patienten und GK fest, dass zwischen beiden Gruppen keine signifikanten hirnstrukturellen Veränderungen bestehen. Diese Beobachtung spiegelt sich in der vorliegenden Studie dahingehend wider, dass die Erkrankung bei den jüngeren Patienten wahrscheinlich nur geringfügig ausgeprägt war und daher auf Grund der hirnstrukturellen Ähnlichkeit zu GK seltener erkannt wurde. Die im Verlauf der Erkrankung progrediente Hirnatrophie (Filippi et al., 2020) erklärt die höhere Vorhersagegenauigkeit des T1-Classifiers bei älteren Probanden. Das Geschlecht schien hingegen keinen wesentlichen Einfluss auf die Korrektheit einer Vorhersage zu haben. Folglich legen die Ergebnisse nahe, dass strukturelle Aufnahmen nach dem verwendeten Vorverarbeitungsverfahren nicht zur Erkennung von IPS im Frühstadium geeignet sind. Der mögliche Einfluss weiterer Confounder wird in Abschnitt 5.5.3 diskutiert.

Die Verwendung der mittels dualer Regression aus den rs-fMRT-Aufnahmen extrahierten Zeitserien zur Klassifizierung mit einem *LSTM*-Modell ist ein neuartiger Ansatz, der nach Literaturrecherche bisher noch nicht für die Identifizierung von Patienten mit IPS angewendet wurde. Ein Abgleich mit der aktuellen Forschungsliteratur zeigte zudem, dass die Zeitserien in den meisten Studien zur Analyse der funktionellen Konnektivität mit nur wenigen Ausnahmen (vgl. Abschnitt 5.2.2) bei der Auswertung der jeweiligen Experimente nicht berücksichtigt wurden (u. a. Bastos und Schoffelen, 2016; Noble et al., 2019). Es ist daher überraschend, dass sich diese innerhalb der HHU-Kohorte für die Differenzierung zwischen Patienten mit IPS und GK als geeignet erwiesen. Im Gegensatz zu den Ergebnissen des T1-Classifiers war die Vorhersagegenauigkeit des rs-fMRT-Classifiers zudem weniger abhängig von Gruppenzugehörigkeit, Alter und Geschlecht (vgl. Tabelle 4.2 und 4.3). Daher ist es wahrscheinlich, dass die Klassifizierung tatsächlich anhand eines IPS-typischen Merkmals erfolgte und nicht von den genannten Confoundern beeinflusst wurde.

Von Bedeutung ist zudem die deutlich präzisere Gruppenvorhersage des rs-fMRT-Classifiers unter Verwendung einer echte *group IC-map* für die duale Regression im Vergleich zur Negativkontrolle, die lediglich die duale Regression mit einer *pseudo group IC-map* umfasste.

Dies ist ein Hinweis darauf, dass der Classifier für die Vorhersage ein Merkmal heranzog, das in engem Zusammenhang mit der funktionellen Konnektivität von Gehirnnetzwerken steht. Weiterhin deutet die gute Differenzierbarkeit der Probanden mit diesen Verfahren darauf hin, dass diese Konnektivität bei Patienten mit IPS auch messbar verändert ist. Dazu ist jedoch anzumerken, dass das zur Erstellung der *pseudo group IC-maps* verwendete weichgezeichnete Rauschen nur bedingt ausreicht, um echte *spatial maps* realistisch nachzubilden (vgl. Abbildung 3.3).

Die Fusion von vier *DTI*-Metriken zu einer einzelnen Karte und deren anschließende Verarbeitung mit einem dreidimensionalen *CNN* stellen nach derzeitigem Kenntnisstand ein Novum dar. Jedoch war die Vorhersagegenauigkeit mit diesem Verfahren im Vergleich zu den anderen angewendeten Verfahren lediglich gleichwertig oder geringer. Da von 29 Probanden aus der HHU-Kohorte nach Matching keine diffusionsgewichteten Aufnahmen akquiriert wurden und außerdem keine entsprechenden Aufnahmen für eine externe Validierung zur Verfügung standen, ist eine abschließende Evaluation des DTI-Classifiers problematisch. Die unzureichende Vorhersagegenauigkeit bei der Klassifizierung vorhandener Aufnahmen lässt jedoch vermuten, dass diese Bildgebungsmodalität für die Erkennung von IPS nicht geeignet ist. Diese Annahme wird dadurch gestützt, dass vergleichsweise wenige Autoren diffusionsgewichtete Aufnahmen im Rahmen von Studien als Differenzierungsgrundlage verwenden und entsprechende Verfahren den alternativen Ansätzen in der Regel unterlegen sind (vgl. Abschnitt 5.2.3).

Das Ziel für die Entwicklung des multimodalen Classifiers bestand darin, die Vorhersagegenauigkeit gegenüber den unimodalen Classifiern durch Berücksichtigung von Alter und Geschlecht der Probanden zu verbessern. Dieses Ziel wurde erreicht, sodass gezeigt werden konnte, dass die zusätzliche Akquisition struktureller sowie diffusionsgewichteter Aufnahmen sinnvoll ist, obwohl diese, isoliert betrachtet, im Vergleich zu rs-fMRT-Aufnahmen weniger geeignet für die Differenzierung zwischen Patienten mit IPS und GK zu sein scheinen. Alter und Geschlecht von Probanden sind typischerweise bereits bekannt und unkompliziert zu erheben. Sie sollten daher bei Modellen für die computerassistierte Diagnostik grundsätzlich berücksichtigt werden. Ob und inwiefern es sinnvoll ist, den multimodalen Classifier im klinischen Alltag zu verwenden, wird in Unterkapitel 5.3 erläutert.

Beide Kalibrierungskurven und der *Brier score*, mit denen die Zuverlässigkeit der berechneten Vorhersagen bewertet wurden, belegen, dass die Klassifizierungsergebnisse des multimodalen Classifiers gut kalibriert sind und daher zur Abschätzung der Wahrscheinlichkeit für das Vorliegen von IPS bei einem Probanden aus der HHU-Kohorte genutzt werden können. Die auf Basis der Trainingsdaten und unter Anwendung von *LR* bzw. *SVM* berechneten Kalibrierungskurven zeigen darüber hinaus, dass bei einer vorhergesagten Wahrscheinlichkeit von unter 40 % die tatsächliche Häufigkeit für IPS niedriger, während bei einer vorhergesagten Wahrscheinlichkeit von über 40 % die tatsächliche Häufigkeit für IPS höher liegt (vgl. Abbildung 3.5 links). Nach der Modellkalibrierung konnte der *Brier score*, der anhand einer unabhängigen, im Rahmen dieser Kalibrierung abgespaltenen Testkohorte berechnet wurde, signifikant reduziert werden. Entgegen den Erwartungen stieg der *Brier score* bei einigen Vorhersagen aus der Trainingskohorte, die zur Entwicklung der neuen Modellkalibrierung eingesetzt wurde, tendenziell an. Eine mögliche Erklärung dafür ist, dass die Kalibrierungskurven, auf deren Grundlage die neue Kalibrierung ermittelt wurde, insbesondere bei einer geringen Anzahl an Datensätzen nur begrenzt zur Darstellung der bestehenden Modellkalibrierung verwendet werden können. Dies ist vor allem dann der Fall, wenn die Klassifizierungsergebnisse des Modells in Form von Prozentangaben nicht gleichmäßig über ein Intervall von null bis 100 % verteilt sind oder eine falsche Anzahl an *bins* zur Erstellung der Kalibrierungskurven

festgelegt wurde (vgl. Zadrozny und Elkan, 2002).

Eine weitere zentrale Erkenntnis der vorliegenden Studie besteht darin, dass die vorgestellten Klassifizierungsverfahren eine unzureichende externe Validität aufweisen. Die Vorhersagegenauigkeit der Modelle, die mit den Daten aus der HHU-Kohorte – und in einigen Fällen auch zusätzlichen mit denen aus externen Kohorten – trainiert wurden, war bei der Validierung an einer anderen externen Kohorte deutlich geringer als bei der internen Validierung. Umgekehrt war das Modell, das mit den strukturellen Aufnahmen aus externen Kohorten trainiert wurde, wenig leistungsstark bei der Klassifizierung von Aufnahmen aus der HHU-Kohorte. Die Ursachen für diese geringe externe Validität sind vielschichtig. Beispielsweise untersuchten Badea et al. (2017) ebenfalls Veränderungen der funktionellen Konnektivität von Patienten mit IPS und GK innerhalb verschiedener Kohorten und stellten fest, dass die innerhalb einer Kohorte aufgedeckten Unterschiede bei anderen Kohorten nicht reproduzierbar waren. Dies war den Autoren zufolge zum einen auf die Heterogenität der Erkrankung zurückzuführen und zum anderen auf technische Unterschiede bei der Bildakquisition. Beide Punkte werden in Abschnitt 5.5.2 diskutiert.

5.2 Vergleich zu anderen Studien

In den folgenden Abschnitten werden die in der vorliegenden Studie vorgestellten Classifier mit den Ansätzen anderer Autoren zur Detektion von Patienten mit IPS verglichen. Neben der eigentlichen Klassifizierungsleistung wird dabei berücksichtigt, an welchen Kohorten diese Ansätze getestet wurden und ob sie methodische Schwächen aufwiesen, z. B. im Hinblick auf die Verteilung von Confoundern.

5.2.1 Strukturelle Aufnahmen

Esmaeilzadeh et al. (2018) gaben an, mit Hilfe eines dreidimensionalen *CNN* bei ausgewählten Probanden (452 Patienten mit IPS, 204 GK) aus der PPMI-Studie anhand von strukturellen MRT-Aufnahmen ohne weitere Angaben zur Sequenz mit einer *accuracy* von 100 % zwischen Patienten mit IPS und GK differenzieren zu können. Obwohl Patienten mit IPS in deren Kohorte deutlich überrepräsentiert waren, schien die Alters- und Geschlechterverteilung in beiden Gruppen ausgeglichen zu sein. Die Vorverarbeitung der Aufnahmen umfasste *brain extraction* sowie eine vertikale Spiegelung im Sinne von *data augmentation* zur Erhöhung der Anzahl an Datensätzen. Das verwendete *CNN* zog neben der eigentlichen Aufnahme auch Angaben zum Alter sowie Geschlecht der Probanden zur Klassifizierung heran, sodass der Ansatz nicht streng unimodal ist. Obwohl die Studie keine offensichtlichen methodischen Mängel aufweist, sollten ihre Ergebnisse kritisch betrachtet werden, da andere Autoren bisher keine vergleichbare *accuracy* mit ähnlichen Methoden erzielen konnten.

Um die bei IPS pathologisch veränderte Substantia nigra präziser beurteilen zu können, verwendeten Shinde et al. (2019) für die Unterscheidung zwischen betroffenen Patienten und GK eine Neuromelanin-sensitive MRT-Sequenz zur Bildakquisition bei einer privaten Kohorte (45 Patienten mit IPS, 35 GK). Es wurde pro Proband jeweils nur ein einzelner axialer *slice* auf Höhe des Hirnstamms untersucht – die weitere Verarbeitung erfolgte unter Verwendung eines zweidimensionalen *CNN*. Damit erzielten die Autoren im Rahmen einer 5-facher Kreuzvalidierung eine *accuracy* von 84 %. Die Alters- und Geschlechterverteilung in IPS- und GK-Gruppe unterschied sich laut den Autoren nicht signifikant voneinander.

Chakraborty et al. (2020) analysierten ebenfalls ausgewählte Aufnahmen (203 Patienten mit IPS, 203 GK) aus der PPMI-Studie. Die Autoren gaben dabei nur die Geschlechterver-

teilung und das Durchschnittsalter aller 406 Probanden an, ohne zwischen IPS- und GK-Gruppe zu differenzieren. Daher kann nicht ausgeschlossen werden, dass die Klassifizierung durch die Confounder Alter und Geschlecht beeinflusst wurde. Als Modell verwendeten sie ein dreidimensionales *CNN* und evaluierten dieses mit 5-facher Kreuzvalidierung, wobei sie eine *accuracy* von 95 % angaben. Die Vorverarbeitung umfasste lediglich die Bildregistrierung der Aufnahmen anhand einer Referenzaufnahme.

Im gleichen Jahr publizierten Sivaranjini und Sujatha (2020) eine Studie, in der sie das Modell AlexNet (Krizhevsky et al., 2012) nutzten, um mittels *transfer learning* bei ausgewählten Probanden (100 Patienten IPS, 82 GK) aus der PPMI-Studie anhand T2-gewichteter Aufnahmen vorhersagen zu können, ob bei diesen IPS vorlag. AlexNet ist ein *CNN*-Modell, das bereits mit über einer Million verschiedener Bildern vortrainiert wurde, sodass bei der Anwendung auf neue Datensätze nur noch geringfügige Korrekturen der Gewichte erforderlich sind. Eine Anpassung der Geschlechterverteilung wurde offenbar nicht vorgenommen, sodass Frauen in der GK-Gruppe im Vergleich zur IPS-Gruppe überrepräsentiert waren. Die Aufnahmen wurden zu zweidimensionalen *slices* zerlegt, normalisiert und mit einem Gauß-Filter weichgezeichnet. In einem *80:20 split* lag die *accuracy* für die Vorhersage der Testdaten bei 89 %, wobei nicht eindeutig klar ist, ob Training und Validierung auf Probanden- oder auf *slice*-Niveau erfolgten.

Ein Jahr später stellten Bhan et al. (2021) ohne weitere Angaben zu Ein- und Ausschlusskriterien eine sehr kleine Subkohorte (30 Patienten mit IPS, 26 GK) im Alter zwischen 60 bis 75 Jahren aus der PPMI-Studie zusammen und zerlegten die dreidimensionalen Aufnahmen der Probanden ebenfalls zu zweidimensionalen *slices*, die im *90:10 split* zum Training und zur Validierung eines zweidimensionalen *CNN*-Modells genutzt wurden. Die Confounder Alter und Geschlecht wurden scheinbar von den Autoren nicht berücksichtigt, zudem bleibt auch bei dieser Studie unklar, ob Training und Validierung auf Probanden- oder auf *slice*-Niveau erfolgten. Insgesamt muss die angegebene Klassifizierungsgenauigkeit von 98 % daher kritisch betrachtet werden. Die genannte Studie weist neben der ungewöhnlich kleinen Stichprobengröße von lediglich 56 Probanden weitere Unstimmigkeiten auf. Beispielsweise wurden mit dem Modell der Autoren in jeder Trainingsepoche die Probanden aus den Testdaten präziser klassifizierte als diejenigen aus den Trainingsdaten. Dieses Phänomen könnte rein zufällig entstanden sein, allerdings sind auch methodische Mängel als mögliche Ursache denkbar.

Abschließend kann festgehalten werden, dass der in der vorliegenden Studie vorgestellte T1-Classifier tendenziell weniger zur Klassifizierung struktureller Aufnahmen geeignet ist als die Verfahren der aufgeführten Autoren. Ein direkter Vergleich ist schwierig, da bei den aufgeführten Ansätzen möglicherweise Confounder nicht berücksichtigt wurden, die Validierung auf *slice*- anstatt auf Probandenniveau erfolgte und auch weitere methodische Ungenauigkeiten nicht ausgeschlossen werden konnten. Auf Grund der heterogenen Präsentation des IPS beeinflusst zudem die Auswahl der Patientenkohorte, inwiefern zwischen Patienten mit IPS und GK differenziert werden kann. Ein fortgeschrittenes Krankheitsstadium erleichtert beispielsweise die Klassifizierung infolge des mit IPS assoziierten verfrühten Einsetzens von leicht nachweisbarer Hirnatrophie, während in einer frühen Krankheitsphase kaum Atrophie festgestellt werden kann (vgl. Abschnitt 5.5.3). Trotzdem erhält der T1-Classifier seine Legitimation durch die Erzeugung von Klassifizierungsergebnissen, die wiederum zu einer Verbesserung der Vorhersagegenauigkeit des multimodalen Classifiers beitragen (vgl. Tabelle 4.5).

5.2.2 Funktionelle Aufnahmen

Abós et al. (2017) verwendeten rs-fMRT-Aufnahmen einer privaten Kohorte (27 Patienten mit IPS und leichter kognitiver Störung, 43 Patienten mit IPS ohne kognitive Störung, 38 GK), um mittels multivariate Analysemethoden (*multivariate pattern analysis*) in Kombination mit *randomized logistic regression* verwendbare *features* zu berechnen, die anschließend zur Klassifizierung mit Hilfe eines *SVM*-Modells genutzt werden konnten. Die Autoren berücksichtigten dabei als Confounder neben der Alters- und Geschlechterverteilung in den Gruppen auch die Händigkeit sowie die Anzahl der von den Probanden absolvierten Schuljahre. Sie erreichten mit ihrer Methode in der *leave-one-out*-Kreuzvalidierung eine *accuracy* von 83 %.

Rubbert et al. (2019) nutzten ebenfalls rs-fMRT-Aufnahmen aus der HHU-Kohorte nach eigenem Matching (42 Patienten mit IPS in Off-Phase, 47 GK) als Grundlage für die Differenzierung. Dabei verglichen sie mittels dualer Regression die Ruhenetzwerke jedes Probanden mit den Ruhenetzwerken bzw. den *group IC-maps* der Probanden zweier externer Kohorten, die ausschließlich aus GK bestanden. Als wesentlichen Unterschied zum rs-fMRT-Classifier berechneten die Autoren daraus unterschiedliche Korrelationskoeffizienten und trainierten damit ein logistisches Regressionsmodell, das im Rahmen der 10-fachen Kreuzvalidierung eine *accuracy* von 76 % bot. Alter und Geschlecht der Probanden unterschieden sich in beiden Gruppen nicht signifikant voneinander.

Garg und McKeown (2019) entwickelten im gleichen Jahr ein *LSTM*-Modell, das getrennt für jede Hirnhemisphäre den Zeitverlauf eines mittels rs-fMRT gemessenen Signals an unterschiedlichen *regions of interest* als Eingabedaten verwendete. Dieses Modell testeten sie anschließend an einer selbst zusammengestellten Kohorte (282 Patienten mit IPS, 92 GK). Im *90:10 split* lag die Sensitivität bei 73 %. Ebenso hoch war der Anteil korrekt als Patienten mit IPS klassifizierter Probanden an der Gesamtheit aller als Patienten mit IPS klassifizierten Probanden.

Ein Jahr später extrahierten Tian et al. (2020) aus den rs-fMRT-Aufnahmen einer privaten Kohorte (72 Patienten mit IPS, 89 GK) die Amplituden niedrigfrequenter Fluktuationen des *BOLD*-Signals in bestimmten Hirnregionen. Dabei differenzierten sie zwischen vier unterschiedlichen Frequenzspektren, anhand derer nach Klassifizierung mit einem *SVM*-Modell in der *leave-one-out*-Kreuzvalidierung eine *accuracy* von jeweils 71, 77, 60 sowie erneut 60 % erreicht wurde. Die IPS- und GK-Gruppe waren in Bezug auf Alter, Geschlecht und Anzahl absolvierter Schuljahre nahezu ideal gematcht.

Shi et al. (2022) verwendeten einen ähnlichen Ansatz, indem sie u. a. aus den Amplituden niedrigfrequenter Fluktuationen in unterschiedlichen *regions of interest* die für eine Klassifizierung relevanten *features* aus den Aufnahmen der Probanden generierten und damit ein *SVM*-Modell trainierten. Neben einer privaten Kohorte (59 Patienten mit IPS, 41 GK) für die interne Validierung nutzten sie zudem die NEUROCON-Kohorte (27 Patienten mit IPS, 16 GK) zur externen Validierung. Im Rahmen der interne Validierung erreichten sie mit 10-facher Kreuzvalidierung eine *balanced accuracy* von 80 %, bei Training mit der eigenen Kohorte und Validierung an der NEUROCON-Kohorte kamen sie jedoch nur auf 66 %. Im Hinblick auf das Alter, das Geschlecht sowie die Anzahl absolvierter Schuljahre unterschieden sich die IPS- und GK-Gruppe in der Trainingskohorte nicht signifikant voneinander.

Erwähnenswert ist außerdem eine Studie von Yan et al. (2019), in der die Autoren ebenfalls die aus den rs-fMRT-Aufnahmen extrahierten Zeitserien verwendeten, um unter Anwendung von sowohl klassischen *machine learning* als auch *deep learning* zwischen Patienten mit Schi-

zophrenie ($n = 558$) und GK ($n = 542$) differenzierten. Als wesentliche Unterschiede zum rs-fMRT-Classifier nutzten sie dabei die Zeitserien aus nur 50 *ICs* und schlugen als Modell eine Kombination aus einem *CNN* und einem *RNN* vor. Darüber hinaus führten sie die *group ICA* mit den Aufnahmen der zu untersuchenden Kohorte durch, anstatt dafür auf eine externe Kohorte zurückzugreifen. Im Rahmen eines *leave-one-site-out*-Ansatzes zur Validierung erzielten sie eine *balanced accuracy* von 81 %. Neben vielen weiteren Varianten testeten sie auch ein *LSTM*-Modell, jedoch ausschließlich in Kombination mit einem *CNN*. Dieses war allerdings dem vorgeschlagenen Modell der Autoren unterlegen. Beide Gruppen waren im Hinblick auf Alter und Geschlecht gematcht.

Auf Grund der verschiedenen Kohorten ist ein direkter Vergleich des rs-fMRT-Classifiers mit den Ansätzen der aufgeführten Autoren praktisch nicht möglich. Eine Ausnahme stellt die Studie von Shi et al. (2022) dar, da die Autoren ihr Modell ebenfalls an der NEUROCON-Kohorte validierten und dieses, bezogen auf die *balanced accuracy*, im Vergleich um 3,8 Prozentpunkte besser abschnitt als das in der vorliegenden Studie vorgestellte Verfahren. Bei reiner Betrachtung der Evaluationsmetriken für die interne Validierung ist der rs-fMRT-Classifier den Ansätzen der hier vorgestellten Autoren jedoch wahrscheinlich überlegen. Zudem ist davon auszugehen, dass dieser bei 10-facher oder *leave-one-out*-Kreuzvalidierung auf Grund der höheren Anzahl an Trainingsdatensätzen noch präziser klassifizieren könnte.

5.2.3 Diffusionsgewichtete Aufnahmen

Haller et al. (2012) unterschieden bei einer privaten Kohorte (17 Patienten mit IPS, 23 Kontrollprobanden mit Parkinsonismus) anhand von *DTI*-Datensätzen zwischen Patienten mit IPS und Probanden mit Parkinsonismus (z. B. Multisystematrophie, Lewy-Body-Demenz oder vaskuläres Parkinson-Syndrom) ohne IPS. Zu diesem Zweck berechneten die Autoren die fraktionale Anisotropie und extrahierten daraus solche Voxel, in denen beide Gruppen statistisch signifikant voneinander variierten. Mit diesen *features* trainierten sie ein *SVM*-Modell und validierten es mittels 10-facher Kreuzvalidierung, wobei sie eine *accuracy* von 98 % erzielten. Obwohl sich das Alter, das Geschlecht sowie die Ausprägung von Leukenzephalopathie in beiden Gruppen nicht signifikant voneinander unterschieden, weist die beschriebene Studie neben der zu geringen Kohortengröße möglicherweise eine weitere methodische Schwäche auf: Falls die gleichen Probanden, die bei der Auswahl der relevanten Voxel berücksichtigt wurden, auch zur Evaluation des *SVM*-Modells herangezogen wurden, wäre ein Datenleck (*data leakage*) auf Grund von *double dipping* zu befürchten. *Double dipping* bezeichnet die unerlaubte Praxis, für das Training und die Bewertung eines Modells denselben Datensatz zu verwenden.

Prasuhn et al. (2020) erprobten die Klassifizierung an einem *DTI*-Datensatz (162 Patienten mit IPS, 70 GK) aus der PPMI-Studie. Sie berechneten zu diesem Zweck u. a. die üblichen Metriken der fraktionalen Anisotropie, der radialen Diffusionsfähigkeit sowie der mittleren und axialen Diffusivität im Bereich der Substantia nigra und trainierten damit ein *SVM*-Modell. Bei einer *balanced accuracy* von jeweils ca. 50 % kamen sie zu dem Schluss, dass eine *DTI*-basierte Analyse der Substantia nigra nicht geeignet sei, um IPS zu erkennen. Als Confounder wurden das Alter, das Geschlecht sowie das intrakraniale Volumen berücksichtigt.

Yasaka et al. (2021) nutzten ein zweidimensionales *CNN*-Modell, um bei einer privaten Kohorte (115 Patienten mit IPS, 115 GK) die Patienten mit IPS zu identifizieren. Dabei verarbeiteten sie unterschiedliche Varianten der diffusionsgewichteten Bildgebung wie *DTI*, Diffusions-Kurtosis-Bildgebung, *neurite orientation dispersion and density imaging* und *g-ratio mapping* jeweils zu einer 60×60-Matrix, die die Eingabedaten für ein *CNN* bildete. Auf diese Weise erreichten sie in der 5-fachen Kreuzvalidierung eine *accuracy* von 67 % unter Verwendung der fraktionalen Anisotropie und 81 % für die Diffusions-Kurtosis-Bildgebung,

während die anderen Bildgebungsmodalitäten im Hinblick auf ihre Eigenschaft zur Korrektklassifikation dazwischen lagen. Das Alter und das Geschlecht der Probanden unterschieden sich in beiden Gruppen nicht signifikant voneinander.

Ein Jahr später testeten Zhao et al. (2022) ein dreidimensionales *CNN*-Modell an einer privaten Kohorte (305 Patienten mit IPS, 227 GK). Dazu berechneten sie die fraktionale Anisotropie sowie die mittlere Diffusivität und teilten danach das Gehirn in 90 *regions of interest* ein. Anschließend trainierten sie ein *CNN* mit den Trainingsdaten aus einem *80:20 split* separat für jede der 90 Hirnregion. Im nächsten Schritt passten Sie ihren Ansatz soweit an, dass Gehirnregionen, die für die Klassifizierung irrelevant waren, nicht weiter berücksichtigt wurden. Die abschließende Evaluation des Modells an den Testdaten ergab eine *accuracy* von 78 %. Bezüglich der Alters- und Geschlechterverteilung machten die Autoren nur Angaben, die sich auf die gesamte Kohorte bezogen. Dadurch ist die Verteilung dieser Confounder in den Trainings- und Testdaten nicht ersichtlich.

Insgesamt wurden diffusionsgewichtete Aufnahmen nur selten zur Detektion von Patienten mit IPS verwendet. Dies liegt u. a. daran, dass deren Vorverarbeitung aufwendig ist und noch keine eindeutigen *features* entdeckt wurden, die eine sichere Differenzierung ermöglichen. Ein direkter Vergleich des DTI-Classifiers mit den Ansätzen der vorgestellten Autoren ist auch hier auf Grund der unterschiedlichen Kohorten nicht möglich.

5.2.4 Multimodale Ansätze

Durch die Kombination unterschiedlicher Untersuchungsmethoden gelang es Prashanth et al. (2016), Patienten mit IPS im Frühstadium mit hoher Genauigkeit von GK zu unterscheiden. Sie nutzten die PPMI-Studie zur Erstellung einer Kohorte (401 Patienten mit IPS, 183 GK) und analysierten sowohl die Leistung der Probanden bei einem Riechtest, als auch deren Antworten auf einem Fragebogen zur Identifizierung von Schlaf-Verhaltensstörungen, Biomarker aus dem Liquor cerebrospinalis sowie die Bindungsverhältnisse von Ioflupane I-123 in einer striatalen Region im Vergleich zu anderen Gehirnregionen, die mittels Einzelphotonen-Emissionscomputertomographie (*single photon emission computed tomography*) ermittelt wurden. Daraus generierten sie 13 *features* und berechneten, dass in beiden Gruppen elf davon für die Differenzierung statistisch signifikant waren. Anschließend trainierten und evaluierten sie mit diesen elf *features* eine Vielzahl an unterschiedlichen konventionellen *machine learning*-Modellen. Die höchste durchschnittliche *accuracy* von 96 % wurde in der 5-fachen Kreuzvalidierung mit einem *SVM*-Modell erreicht. Kritisch ist dabei zu betrachten, dass die Autoren lediglich die Alters-, jedoch nicht die Geschlechterverteilung in den Gruppen berücksichtigten. Zudem könnten *double dipping* und folglich *data leakage* aufgetreten sein, falls die *features* anhand der gleichen Datenmenge ausgewählt wurden, die später der Evaluierung des Modells diente.

Vásquez-Correa et al. (2018) analysierten die Gangart, die Handschrift und die Sprache einer privaten Kohorte (44 Patienten mit IPS, 40 GK) mit dem Ziel, zwischen Patienten mit IPS und GK zu differenzieren. Bei allen Probanden wurde insbesondere die Motorik zu Beginn und Ende des Geh-, des Schreib- und des Sprechvorgangs mit einem *CNN* untersucht. Die Autoren stellten fest, dass ein *CNN* bei Verwendung aller drei Untersuchungsmodalitäten präziser differenzieren konnte als ein Modell, das nur eine einzelne Untersuchungsmodalität nutzte. Bei der multimodalen Analyse der Parameter zu Beginn der jeweiligen Bewegung erzielten sie im *90:10 split* bei den Testdaten eine *accuracy* von 98 %. Obwohl die Verteilung von Alter und Geschlecht für die gesamte Kohorte angegeben wurde, ist die Zusammensetzung der Probanden nach Aufteilung in Trainings- und Testdaten nicht bekannt. Die Testdaten enthielten maximal neun Probanden, sodass keine Verallgemeinerung der Ergebnisse möglich ist.

Talai et al. (2021) nutzten die Kombination T1- und T2-gewichteter Aufnahmen sowie *DTI* von Probanden einer privaten Kohorte des Universitätsklinikums Hamburg-Eppendorf, um zwischen Patienten mit IPS ($n = 45$), solchen mit progressiver supranukleärer Blickparese ($n = 20$) und GK ($n = 38$) zu unterscheiden. Nach Vorverarbeitung der Aufnahmen wurden je nach Bildgebungsmodalität 234, 396 bzw. 520 *features* extrahiert und nach ihrer Relevanz zur Differenzierung bewertet. Anschließend wurden die 100 besten *features* genutzt, um eine Vielzahl an unterschiedlichen konventionellen *machine learning*-Verfahren zu trainieren und zu evaluieren. Die höchste *accuracy* von 95 % wurde in der *leave-one-out*-Kreuzvalidierung bei Kombination eines *SVM*-Modells mit einem *multilayer perceptron* erreicht. Die Autoren wiesen auf die unterschiedliche Alters- und Geschlechterverteilung in den einzelnen Gruppen hin. Ein mögliches *double dipping* bei Bestimmung der zur Differenzierung sinnvollsten *features* und der anschließenden Modellevaluierung schlossen sie jedoch aus. Im Gegensatz zu Vásquez-Correa et al. (2018) stellten sie bei der Verwendung von drei Bildgebungsmodalitäten keinen messbaren Vorteil gegenüber der ausschließlichen Verwendung der *DTI*-Datensätze fest.

Obwohl die vorgestellten Studien auf Grund zu kleiner Kohorten, der Nichtberücksichtigung von Confoundern oder des potentiellen *double dipping* Schwächen aufweisen, sind die Korrektklassifikationsraten der Modelle dennoch beachtlich. Bei Verwendung multimodaler Eingabedaten war die Vorhersagegenauigkeit der vorgestellten Modelle mindestens gleichwertig zu Modellen mit Verwendung der besten Einzelbildgebungsmodalität. Es fällt jedoch auf, dass in keiner der vorgestellten Studien — trotz der Möglichkeit, die Daten unkompliziert zu erheben — das Alter oder das Geschlecht der Probanden als weitere *features* einbezogen wurden. Diese Angaben sind zwar nicht dazu geeignet, zwischen Patienten mit IPS und GK zu unterscheiden, jedoch unterstützen sie das Modell bei der Bewertung bzw. Relativierung weiterer Merkmale, die sich je nach Alter oder Geschlecht unterscheiden. Neurophysiologisch relevante Beispiele solcher Merkmale umfassen umfassen das Kortexvolumen (Peters und Morrison, 2012) und das Bindungsverhalten striataler Dopaminrezeptoren (Pohjalainen et al., 1998). Wie in der vorliegender Studie gezeigt wurde, verbesserte sich die *balanced accuracy* des multimodalen Classifiers allein durch die Berücksichtigung von Alter und Geschlecht der Probanden um ungefähr einen Prozentpunkt (vgl. Tabelle 4.5).

5.3 Klinische Anwendung

Eine Beurteilung, ob die vorgestellten Verfahren als Biomarker für die Diagnostik des IPS verwendet werden können, erfolgt in der Regel anhand der Parameter Relevanz und Validität (Strimbu und Tavel, 2010). Unter Berücksichtigung dieser Parameter wird abschließend diskutiert, ob eine klinische Anwendung der Verfahren als Screening-Untersuchung sinnvoll ist.

5.3.1 Relevanz

Der Parameter Relevanz bezieht sich auf die Eigenschaft eines Biomarkers, klinisch relevante Fragen bei der Diagnostik zu beantworten. Eine häufige Ursache für die Fehldiagnose einer Erkrankung als IPS liegt darin, dass andere Erkrankungen mit Parkinsonismus, wie der essentielle Tremor, das vaskuläre Parkinson-Syndrom, die Lewy-Body-Demenz, die progressive supranukleäre Blickparese oder die Multisystematrophie, eine ähnliche Symptomatik wie IPS verursachen und daher ausgeschlossen werden müssen (Tolosa et al., 2006). Für die Bewertung der Relevanz ist zu beachten, dass die in der vorliegenden Studie vorgestellten Classifier zunächst mit dem Ziel entwickelt wurden, lediglich zwischen Patienten mit IPS und GK ohne

Parkinsonismus zu differenzieren, um auf diese Weise ein Fundament für die weitere systematische Entwicklung eines Biomarkers zu schaffen. Darauf aufbauend kann in zukünftigen Studien die Trainingskohorte um Probanden mit Parkinsonismus erweitert werden, um die Relevanz der Classifier für den klinischen Alltag zu steigern.

Auf Grund der unspezifischen Symptome von IPS in einem frühen Stadium ist die sichere Früherkennung der Erkrankung eine weitere klinische Herausforderung mit hoher Relevanz (Jankovic et al., 2000). Von den 58 Patienten mit IPS aus der HHU-Kohorte nach Matching, bei denen das Stadium nach Hoehn und Yahr erhoben wurde, sind vier Probanden dem Stadium I und 27 Probanden dem Stadium II zuzuordnen. Diese Stadien entsprechen einer einseitigen (Stadium I) bzw. einer beidseitigen (Stadium II) Bewegungsstörung ohne Haltungsinstabilität. Asymptomatische Patienten wurden nicht in die HHU-Kohorte eingeschlossen. Das bedeutet, dass die Classifier grundsätzlich dazu geeignet sind, bei symptomatischen Patienten IPS zu erkennen. Die Klassifizierungsleistung bei Anwendung an asymptomatischen Patienten ist hingegen unklar.

Zusammenfassend kann festgestellt werden, dass die Classifier in ihrem derzeitigen Entwicklungsstand im Hinblick auf den Parameter Relevanz in einem klinischen Setting noch nicht ausreichend beurteilt werden können. Dies ist vor allem darauf zurückzuführen, dass ausschließlich Unterschiede zwischen Patienten mit IPS und GK untersucht wurden und Probanden mit Parkinsonismus ohne IPS nicht Bestandteil der herangezogenen Kohorten waren. Mit Hilfe von Datensätze dieser Patientengruppe ist es jedoch voraussichtlich unproblematisch, die Modelle entsprechend zu trainieren und neu zu evaluieren.

5.3.2 Validität

Der Parameter Validität bezieht sich auf die Eigenschaft eines Biomarkers, klinische Endpunkte zu charakterisieren. In der vorliegenden Studie entspricht dies der Unterscheidung zwischen Patienten mit IPS und GK. Die interne und externe Validität der verwendeten Verfahren sind in den vorangegangenen Unterkapiteln 4.1, 4.2 und 4.3 aufgeführt. Neben der hohen internen Validität fiel vor allem deren große Diskrepanz zur externen Validität auf. Des Weiteren konnte an den Kalibrierungskurven abgelesen werden, dass die vom multimodalen Classifier vorhergesagte Wahrscheinlichkeit dafür, dass bei einem Probanden aus der HHU-Kohorte nach Matching IPS vorlag, mit der tatsächlichen Wahrscheinlichkeit dafür meist übereinstimmte.

Die ungenügende externe Validität spricht gegenwärtig gegen die Verwendung der Verfahren als Biomarker zur Diagnostik des IPS außerhalb von Probanden der HHU. Ein möglicher Ansatz zur Überwindung dieser Einschränkung besteht darin, die Trainingskohorten mit Datensätze aus externen Bildgebungsstudien zu erweitern und die Modelle damit neu zu trainieren sowie zu evaluieren.

5.3.3 Verwendung als Screening-Biomarker

Im Folgenden wird überprüft, ob die unter Anwendung des multimodalen Classifiers erzielte Sensitivität und Spezifität ausreichend ist, um dieses Verfahren im Rahmen einer Screening-Untersuchung zur Erkennung von IPS einzusetzen.

Eine Screening-Untersuchung dient zur Früherkennung von Erkrankungen bei einer größeren, asymptomatischen Population und umfasst normalerweise einen möglichst sensitiven Suchtest sowie anschließend einen möglichst spezifischen Bestätigungstest für alle zuvor Test-Positiven (Spix und Blettner, 2012). Zur Vereinfachung wird im Folgenden angenommen, dass

die gesamte Bevölkerung im Alter von über 60 Jahren ohne bekannte Vorerkrankungen des zentralen Nervensystems gescreent wird und die Prävalenz von IPS innerhalb dieser Population 1 % beträgt. Des Weiteren wird angenommen, dass das Screening mit dem multimodalen Classifier aus dem *80:20 split* durchgeführt wird und dieser asysmptomatische Patienten mit IPS ähnlich gut wie jene aus der HHU-Kohorte nach Matching detektiert. Demnach liegt die Sensitivität bei 89,2 % und die Spezifität bei 90,0 %.

Zur Einschätzung der Wahrscheinlichkeit, dass bei einer Testperson mit einem positiven Testergebnis auch IPS vorliegt bzw. bei einem negativen Testergebnis kein IPS vorliegt, sind vor allem die entsprechenden prädiktiven Werte relevant (Hoyer und Zapf, 2021). Diese geben unter Berücksichtigung der Prävalenz einer Erkrankung Aufschluss darüber, wie zuverlässig eine bestimmte Vorhersage ist. Der positive prädiktive Wert (*positive predictive value*, *PPV*) repräsentiert demnach die Wahrscheinlichkeit, an IPS erkrankt zu sein, wenn dies vom multimodalen Classifier vorhergesagt wurde. Er wird folgendermaßen berechnet:

$$\text{PPV} = \frac{\text{Sensitivität} \cdot \text{Prävalenz}}{\text{Sensitivität} \cdot \text{Prävalenz} + (1 - \text{Spezifität}) \cdot (1 - \text{Prävalenz})}$$

Das Einsetzen der Zahlenwerte ergibt einen *PPV* von 8,3 %. Im Screening-Setting bedeutet dies, dass die Wahrscheinlichkeit, dass die Test-Positiven tatsächlich an IPS leiden, nur bei 8,3 % liegt.

Mit dem negativen prädiktiven Wert (*negative predictive value*, *NPV*) kann hingegen abgeschätzt werden, wie hoch die Wahrscheinlichkeit ist, dass eine Testperson tatsächlich gesund ist, wenn dies vom multimodalen Classifier vorhergesagt wurde. Die Berechnung des *NPV* erfolgt mit folgender Formel:

$$\text{NPV} = \frac{\text{Spezifität} \cdot (1 - \text{Prävalenz})}{\text{Spezifität} \cdot (1 - \text{Prävalenz}) + (1 - \text{Sensitivität}) \cdot \text{Prävalenz}}$$

Damit ergibt sich ein *NPV* von 99,9 %, sodass bei einem Test-Negativen mit hoher Sicherheit kein IPS vorliegt.

Abschließend kann festgehalten werden, dass der klinische Einsatz der Classifier bei derzeitigem Entwicklungsstand als Biomarker zur Diagnostik des IPS verfrüht wäre. So wurden Patienten ohne IPS, die jedoch andere Erkrankungen des zentralen Nervensystems aufwiesen, in den Trainings- und Testdaten nicht berücksichtigt, sodass nicht abgeschätzt werden kann, welche Vorhersagen die Classifier bei dieser Gruppe ausgeben würden. Des Weiteren ist eine Verwendung der vorgestellten Verfahren auf Grund der geringen externen Validität lediglich bei Datensätzen aus der HHU sinnvoll. Für eine Verwendung im Screening ist die Sensitivität der Classifier für die vergleichsweise niedrige Prävalenz von IPS nicht hoch genug.

5.4 Lösung der Black Box-Problematik

Die Akzeptanz von *machine learning* bei der computerassistierten Diagnostik seitens der Patienten und Ärzte kann erhöht werden, wenn die durch die Software getroffenen Entscheidungen für den Anwender nachvollziehbar sind (Ribeiro et al., 2016). Es ist folglich problematisch, dass selbst Experten für KI zahlreiche Vorhersagen von prädiktiven Modellen nicht erklären können, da sie die inneren Abläufe auf Grund deren Komplexität nicht vollständig verstehen (Carvalho et al., 2019). Zur Verbildlichung bezeichnen einige Autoren solche Modelle daher als *black boxes* (Castelvecchi, 2016; Lei et al., 2018), da die Verarbeitungsschritte der Eingabedaten zu einem Ergebnis nicht ohne größeren Aufwand transparent darstellbar sind. In der vorliegenden Studie wurde daher für den T1-Classifier eine Methode verwendet, die die zu

einer IPS-Klassifizierung geführten Bereiche der MRT-Aufnahme mittels *gradient-weighted class activation mapping* visualisiert. Ein ähnlicher Ansatz wäre für den rs-fMRT-Classifier nicht zielführend, da die als Eingabedaten verwendeten Zeitserien keine Schnittbilder repräsentieren. Auch die Visualisierung der relevanten Bereiche in den diffusionsgewichteten Aufnahmen erscheint wenig sinnvoll, da die Gruppenzugehörigkeit im Vergleich zu den anderen Bildgebungsmodalitäten mit Hilfe des DTI-Classifiers im *80:20 split* nur vergleichsweise selten korrekt vorhergesagt wurde. Es ist daher unwahrscheinlich, dass diese Aufnahmen eindeutige Merkmale enthalten, die zur Identifizierung von IPS dienen könnten.

Für diese Methode wurden die Modelle im *80:20 split* der *hold-out validation* verwendet, da diese eine höhere Klassifikationsgenauigkeit und Reproduzierbarkeit als die Modelle im Rahmen der 5-fachen Kreuzvalidierung boten (vgl. Tabelle 4.4). Das *CNN* des T1-Classifiers wurde mit den Aufnahmen der Probanden aus der HHU-Kohorte nach Matching (vgl. Abschnitt 3.1.1) trainiert und anschließend dafür genutzt, die Aufnahmen der Patienten mit IPS aus den Testdaten zu klassifizieren. Anschließend wurden die *heatmaps* der Probanden, bei denen das *CNN* eine besonders hohe Wahrscheinlichkeit für IPS vorhergesagt hat, weiter untersucht (vgl. Abbildung 4.1). Dabei fiel auf, dass das Modell fast ausschließlich die Bildinformationen aus der Mediansagittalebene unter Aussparung der lateralen Ventrikel zur Klassifizierung verwendete. Des Weiteren zog es bei einigen Probanden den rechten medialen Gyrus temporalis inferior für die Vorhersage heran. Ibarretxe-Bilbao et al. (2011) entdeckten damit übereinstimmend, dass bei Patienten mit IPS die Atrophie des medialen Temporallappens mit dem Ausmaß an Gedächtnisstörungen korreliert. Die extrazerebral hervorgehobenen Bereiche der *heatmap* können als Zeichen dafür gewertet werden, dass das Modell keine sicheren *features* zur Differenzierung innerhalb des Gehirns finden konnte und daher weitere Bereiche der Aufnahme danach absuchte. Nach visuellem Vergleich mit den Aufnahmen von GK konnten keine für eine radiologische Blickdiagnose geeigneten eindeutigen Merkmale für die Differenzierung ermittelt werden. Dies ist jedoch nicht überraschend, da der T1-Classifier mit einer durchschnittlichen *ROC AUC* von 0,81 auch nicht sicher bei der Klassifizierung war.

Zusammenfassend konnten trotz Visualisierung mit *heatmaps* keine für den Anwender nachvollziehbaren Merkmale gefunden werden, die für die Differenzierung zwischen Patienten mit IPS und GK genutzt werden können. Da der T1-Classifier in der 5-fachen Kreuzvalidierung mit unterschiedlich zusammengestellten Trainingsdaten und generell mit zufälligen initialen Gewichten trainiert wurde, ist außerdem zu berücksichtigen, dass die hier vorgestellten *heatmaps* möglicherweise nicht für die übrigen Vorhersagen der Patienten mit IPS aus der HHU-Kohorte repräsentativ sind.

5.5 Limitationen

Die Generalisierbarkeit der Ergebnisse in der vorliegenden Studie ist aus mehreren Gründen eingeschränkt. Die folgenden Abschnitte umfassen eine Übersicht über die wesentlichen Limitationen und methodischen Probleme, die während der Entwicklung der Classifier auftraten und bei deren Evaluierung zu berücksichtigen sind.

5.5.1 Stichprobengröße

Die Anzahl an Datensätzen in den Trainingsdaten und somit auch die Anzahl der in die Studie eingeschlossenen Probanden gelten als zentrale Größe für die Fähigkeit zur Mustererkennung nahezu aller *machine learning*-Modelle (Raudys und Jain, 1991). Auf Grund der relativ hohen Anzahl an *features* stellt die Analyse drei- und vierdimensionaler MRT-Aufnahmen eine

besondere Herausforderung dar, da das Modell neben der Verarbeitung bestimmter *features* auch lernen muss, welche *features* für die Klassifizierung überhaupt relevant sind (Blum und Langley, 1997). Folglich steigt mit der Komplexität der Datensätze deren benötigte Anzahl, um *overfitting* zu verhindern. Unter der Annahme, dass ein Voxel einem *feature* entspricht, ergibt sich für die drei Bildgebungsmodalitäten folgende Verteilung (vgl. Tabelle 3.2 und 3.3): Die strukturellen T1-gewichteten Aufnahmen umfassen ca. 1,1 Millionen *features* bei 135 Datensätzen, die *snippets* aus den rs-fMRT-Aufnahmen haben lediglich 1000 *features* bei 4050 Datensätzen und die diffusionsgewichteten Aufnahmen beinhalten ca. 12,7 Millionen *features* bei nur 106 Datensätzen. Abgesehen von den rs-fMRT-Aufnahmen liegt somit ein starkes Ungleichgewicht zwischen der Anzahl an *features* zu jener der Datensätze vor, was sich in der vergleichsweise niedrigen Klassifizierungsgenauigkeit in diesen Bildgebungsmodalitäten widerspiegelt. Zur Überwindung der Problematik und des daraus resultierenden *overfitting* müssten mehrere Hundert weiterer Probanden rekrutiert werden, was mit einem hohen personellen und finanziellen Aufwand verbunden ist.

Ein ähnliches Phänomen zeigte sich im *leave-one-site-out*-Ansatz: Die Vorhersagegenauigkeit des Classifiers war insbesondere bei der Klassifizierung der Probanden aus der HHU-Kohorte niedrig, da aus diesen im Vergleich zu den anderen Kohorten die meisten Datensätze gebildet werden konnten und somit vergleichsweise wenige Datensätze für das Training zur Verfügung standen. Der umgekehrte Effekt dieses Phänomens zeigte sich darin, dass die Klassifizierung der Probanden aus der Köln-Kohorte gemessen an der *balanced accuracy* am genauesten war, was sich entsprechend damit begründen ließ, dass diese Kohorte die geringste Anzahl an Probanden umfasste und folglich entsprechend viele Datensätze für das Training genutzt werden konnten.

5.5.2 Site Effects

Eine weitere wesentliche Einschränkung der entwickelten Classifier zeigte sich in der geringen Vorhersagegenauigkeit bei deren Anwendung auf den Datensätzen aus unterschiedlichen Kohorten. Dieses Phänomen ist eine möglich Folge von *site effects* bzw. *scanner effects*, die darauf beruhen, dass einzelne Kohorten meist nur eine kleine, homogene Probandengruppe enthalten (Nunes et al., 2020) bzw. die Bildakquisition für jede Kohorte mit kohortenspezifischen oder uneinheitlichen Erfassungsprotokollen erfolgte (Saponaro et al., 2022). Somit haben Ergebnisse, die unter Verwendung von Datensätzen einer einzelnen Kohorte erzielt wurden, keine übertragbare Gültigkeit auf andere Kohorten. Homogene Kohorten sind dahingehend problematisch, dass sie eine heterogene Gesamtbevölkerung, und somit auch andere Kohorten, nur eingeschränkt abbilden. Neurobiologische Merkmale, die zur Beurteilung der Homogenität herangezogen werden können, umfassen u. a. die Ausprägung der untersuchten Erkrankung, das Alter, das Geschlecht, die Ethnizität, die aktuelle Medikation oder vorausgegangene Lebensereignisse der Probanden (Bayer et al., 2022). Diese Merkmale haben potentiell Einfluss auf die Struktur sowie die Funktionsweise des Gehirns und somit auch darauf, wie sich IPS in entsprechenden MRT-Aufnahmen darstellt. Bezüglich *scanner effects* konnte zudem gezeigt werden, dass auch bei einheitlichen Erfassungsprotokollen eine relevante Variabilität zwischen den Aufnahmen unterschiedlicher Scanner besteht (Jovicich et al., 2016), wodurch ein Vergleich von Probanden aus unterschiedlichen Kohorten zusätzlich erschwert wird.

Besonders deutlich wurden diese Phänomene in der vorliegenden Studie bei der externen Validierung des rs-fMRT-Classifiers, da die hohe Klassifizierungsgenauigkeit aus der internen Validierung nicht reproduziert werden konnte. Dies kann u. a. damit begründet werden, dass die relevanten *BOLD*-Signale in rs-fMRT-Aufnahmen grundsätzlich von Störsignalen (vgl. Abschnitt 2.3.8) überlagert sind, sodass vor allem bei dieser Bildgebungsmodalität ein hohes

(zeitliches) Signal-Rausch-Verhältnis besteht (Welvaert und Rosseel, 2013). Eine ähnliche Beobachtung machten Shi et al. (2022) bei der Entwicklung eines auf rs-fMRT-Aufnahmen basierenden Biomarkers zur Erkennung von Patienten mit IPS (vgl. Abschnitt 5.2.2). Während die Autoren für die interne Validierung mit ihrem Modell eine *balanced accuracy* von 80 % erzielten, nahm diese im Rahmen der externen Validierung an der NEUROCON-Kohorte auf nur noch 66 % ab.

Kelly et al. (2022) zeigten in einer systematischen Übersichtsarbeit anhand von 535 Studien zur Anwendung von KI in der Radiologie, dass die Vorhersagegenauigkeit der untersuchten Modelle bei deren externen Validierung in einem Bereich von 4 bis 44 % gegenüber der internen Validierung abnahm. Eine Möglichkeit zur Überwindung dieser Problematik besteht gemäß den Autoren darin, die Bildakquisition an möglichst baugleichen MRT-Scannern mit einheitlichen Erfassungsprotokollen durchzuführen und die Trainingskohorte so zusammenzustellen, dass sie im Hinblick auf die zuvor beschriebenen Merkmale möglichst heterogen ist. Diese Heterogenität kann beispielsweise durch eine Vergrößerung des Einzugsbereichs der Probanden sowie eine Anpassung der Ein- sowie Ausschlusskriterien zugunsten einer Erhöhung der Anzahl potentieller Studienteilnehmer erhöht werden.

5.5.3 Patientenseitige Confounder

Bei der Evaluierung der Classifier ist weiterhin zu beachten, dass sich die Berücksichtigung patientenseitiger Confounder kompliziert gestaltete. Bekannte Confounder bei der Untersuchung von funktioneller Konnektivität umfassen neben den in der vorliegenden Studie kontrollierten Faktoren Alter und Geschlecht auch die Händigkeit (Pool et al., 2015), die Bewegungen während der Bildakquisition (Power et al., 2012), die Größe des Gehirns (Hänggi et al., 2014) und die fluide Intelligenz (Cole et al., 2012). Mögliche Korrelationen dieser Faktoren zu neuronalen Veränderungen bei Patienten mit IPS sind zwar durchaus vorstellbar, jedoch retrospektiv nicht zu kontrollieren, da sie bei den Probanden aus der HHU-Kohorte nicht erhoben wurden. Vor allem bei Patienten mit IPS sollte berücksichtigt werden, dass auch ein Ruhetremor während der Bildakquisition zu Bewegungen des Kopfes führen kann. Die daraus resultierenden Bewegungsartefakte können zu einer inkorrekten Berechnung der funktionellen Konnektivität betroffener Probanden führen (vgl. Abschnitt 2.3.5).

Da vor allem Patienten mit IPS unter 60 Jahren oft fälschlicherweise als GK klassifiziert wurden, kann davon ausgegangen werden, dass ein junges Probandenalter von den Classifiern trotz des verwendeten Matching-Verfahrens als Surrogatmarker für die Zuordnung zu einen GK genutzt wurde. Dieses Phänomen fällt vor allem bei dem T1-Classifier auf und lässt sich mit der physiologische Atrophie des Gehirns mit zunehmendem Alter erklären, die sich insbesondere mit Hilfe struktureller MRT-Aufnahmen darstellen lässt (MacDonald und Pike, 2021). Bei Patienten mit IPS ist der Prozess dieser zunehmenden Hirnatrophie beschleunigt (Filippi et al., 2020), sodass der T1-Classifier nach dem Training eine Hirnatrophie als einen Hinweis auf IPS fehldeuten könnte, während ihr Fehlen auf einen GK hinweise.

5.5.4 Ground Truth

Obwohl die Diagnose des IPS von erfahrenen Ärzten gestellt wurde, erfolgte auf Grund der unverhältnismäßig hohen Invasivität bei keinem Probanden eine histopathologische Sicherung der Erkrankung. Daher besteht ein geringes Maß an diagnostischer Unsicherheit, die sich negativ auf Training und Evaluation der Modelle auswirken kann. Die Unsicherheit über eine korrekte Klassifizierung innerhalb der Datensätze (Grundwahrheit, *ground truth*) ist generell ein Faktor, der bei allen *machine learning*-Modellen zur medizinischen Diagnostik kritisch betrachtet werden sollte (Lebovitz et al., 2021). Speziell für IPS stellten Rizzo et al. (2016)

in einer Metaanalyse fest, dass die Erkrankung nur in 80 % aller Fälle korrekt diagnostiziert wurde. Demnach ist die unzureichende Vorhersagegenauigkeit der Classifier bei den externen Kohorten auch zu einem geringen Anteil durch Fehldiagnosen erklärbar. Durch eine Bereinigung dieser Fehldiagnosen könnte möglicherweise auch die bereits hohe Vorhersagegenauigkeit der Classifier bei Anwendung innerhalb der HHU-Kohorte weiter gesteigert werden.

5.5.5 Softwaretechnische Herausforderungen

Bereits nach der Vorverarbeitung mit dem Softwarepaket FSL konnten durch manuelles Betrachten der Aufnahmen erhebliche Schwankungen in der Qualität der durchgeführten Anpassungen festgestellt werden. Insbesondere *brain extraction* mit FSL BET sowie die Bildregistrierung mit FSL FLIRT und FNIRT führte zu inkonsistenten Ergebnissen (vgl. Abbildung 5.1), sodass vor allem bei der Verarbeitung externer Aufnahmen eine Anpassung der Einstellungsparameter erforderlich bzw. selbst damit keine befriedigende Vorverarbeitung dieser Aufnahmen möglich gewesen wäre. Für einen Vergleich von rs-fMRT-Aufnahmen untereinander ist eine präzise Bildregistrierung von grundlegender Bedeutung, da sie die anatomische Variabilität der zu untersuchenden Probanden reduziert (Ardekani et al., 2004), sodass der Fokus auf die funktionelle Variabilität der Gehirne gelegt werden kann. Somit ist davon auszugehen, dass die Vorhersagegenauigkeit des rs-fMRT- sowie des DTI-Classifiers bei einer präziseren Vorverarbeitung zunehmen könnte.

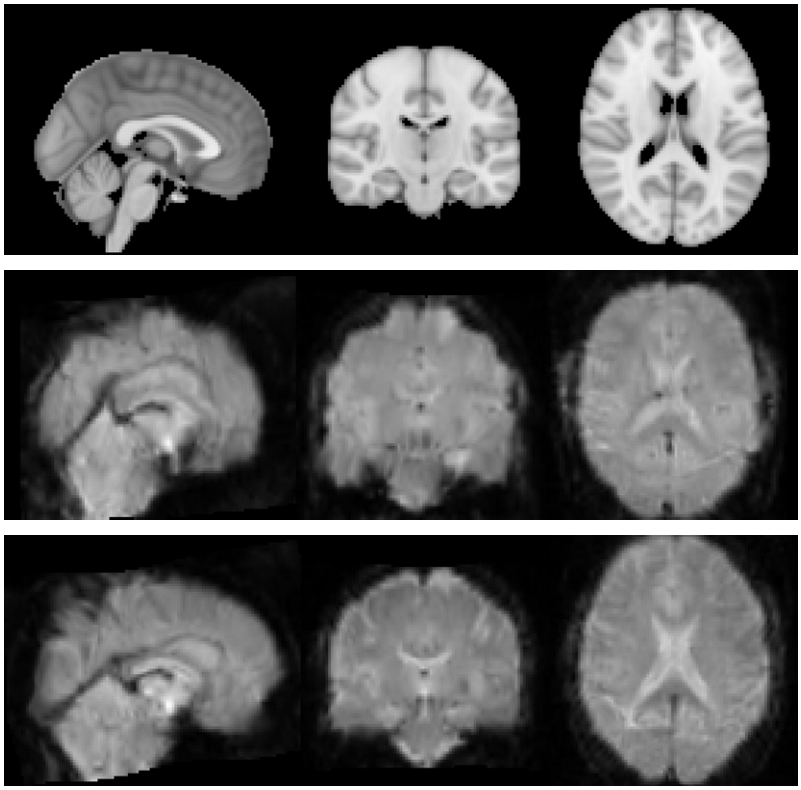

Abb. 5.1: **Vergleich einer Referenzaufnahme mit zwei realen rs-fMRT-Aufnahmen aus der NEUROCON-Kohorte nach Anwendung von FSL FLIRT und FNIRT.** Die erste Zeile zeigt eine Referenzaufnahme (MNI152, vgl. Abschnitt 2.3.3), die das Gehirn anatomisch korrekt darstellt. Die zweite und dritte Zeile zeigen zum Vergleich die *EPI-volumes* zweier Probanden, bei denen das jeweils abgebildete Gehirn durch die Vorverarbeitung mit FSL FLIRT und FNIRT verzerrt wurde.

Ein generelles Problem bei dem Training von *deep learning*-Modellen besteht darin, dass die initialen Gewichte aller trainierbaren *layers* zufällig festgelegt werden und damit Kon-

stellationen auftreten, in denen ihre Adjustierung nicht zur gewünschten Generalisierbarkeit führt (Alahmari et al., 2020). Dieses Phänomen äußerte sich in der vorliegenden Studie darin, dass Evaluationsmetriken wie die *ROC AUC* oder die *balanced accuracy* innerhalb mehrerer Trainingsdurchläufe schwankten und folglich auch die Reproduzierbarkeit der Ergebnisse leicht eingeschränkt war. Ein möglicher Ansatz, um diese Problematik zu beheben, ist der Einsatz von *transfer learning*, das bei vielen medizinischen *deep learning*-Modellen bereits die Vorhersagegenauigkeit verbessern konnte (Yari et al., 2020; Wang et al., 2021). Zu diesem Zweck wird das entsprechende Modell an einem umfangreichen externen Datensatz vortrainiert und der eigentliche Datensatz dient lediglich dem Feintuning der *weights*. Alternativ bleiben bei dem *feature extraction*-Ansatz die *weights* fast aller *layers* nach dem Vortraining unverändert, während durch das Training mit den eigentlichen Datensätzen nur die entsprechenden Parameter der hinteren *layers* auf die neuen Datensätzen angepasst werden (Z. Li und Hoiem, 2017). Beide Verfahren führen dazu, dass die initialen Gewichte der Modelle bereits zu Trainingsbeginn deutlich besser für die eigentlichen Datensätze abgestimmt sind.

5.6 Schlussfolgerungen

In der vorliegenden Studie wurden unterschiedliche Verfahren vorgestellt, in denen strukturelle, funktionelle und diffusionsgewichtete MRT-Aufnahmen sowie Angaben zum Alter und Geschlecht der Probanden verwendet wurden, um zwischen Patienten mit IPS und GK differenzieren zu können. Bei isolierter Betrachtung der genannten Bildgebungsmodalitäten konnte mittels Analyse der rs-fMRT-Aufnahmen die höchste Vorhersagegenauigkeit erzielt werden. Diese ließ sich noch weiter steigern, indem zusätzlich die Klassifizierungsergebnisse der anderen Bildgebungsmodalitäten in Kombination mit den Angaben zum Alter und Geschlecht der Probanden von einem multimodalen Classifier berücksichtigt wurden. Obwohl der multimodale Classifier eine hohe interne Validität aufwies, war die externe Validität im Vergleich dazu deutlich geringer, was u. a. mit *site effects* bzw. *scanner effects* erklärt werden kann. Diese stellen ein häufiges Problem bei bildgebenden Biomarkern dar und können meist nur dadurch behoben werden, dass die Trainingskohorte um eine große Anzahl an externen Datensätzen erweitert wird.

Die vorgestellten Verfahren haben das Potential, die IPS-Diagnostik um einen neuen Biomarker zu bereichern, jedoch bedarf es weiterer Forschung und Optimierung, bevor ein klinischer Einsatz empfohlen werden kann. Dies liegt neben der unzureichenden externen Validität vor allem daran, dass die Verfahren noch nicht an Patienten mit anderen Bewegungsstörungen oder Erkrankungen als IPS getestet wurden. Für eine größere Akzeptanz der Verfahren im klinischen Alltag ist es zudem förderlich, wenn die getroffenen Vorhersagen für den Anwender nachvollziehbar sind. Diese Herausforderung kann bei der Klassifizierung struktureller Aufnahmen z. B. durch die Verwendung von *heatmaps* gelöst werden.

5.7 Ausblick

Dem breiten Angebot an Bildgebungsdaten von Patienten mit IPS und der rasanten Entwicklung sowie Zugänglichkeit von *deep learning* steht ein stetig steigender Bedarf an objektiven Biomarkern für die Diagnostik dieser Erkrankung gegenüber. Daher ist davon auszugehen, dass entsprechende Verfahren künftig weiter verbessert und neue Methoden entwickelt werden, um diese schließlich in der Praxis anwenden zu können. Dieser Trend zeigt sich bereits daran, dass die amerikanische Behörde für Lebens- und Arzneimittel (Food and Drug Administration) im Verlauf der letzten Jahre zunehmend KI-basierte Medizinprodukte zugelassen hat. So standen zu Beginn des Jahres 2023 allein für den Fachbereich Radiologie 190 Produkte dieser Art zur Verfügung (Milam und Koo, 2023). Obwohl sich deren Einsatz in der

Regel auf Routineaufgaben beschränkt, wie etwa dem Erkennen von Lungenrundschatten, ist zu erwarten, dass sie in Zukunft vermehrt in der Diagnostik komplexer Erkrankungen zum Einsatz kommen werden.

Um diesem Bedarf gerecht zu werden, ergeben sich verschiedene Schritte, die auf den Ergebnissen der vorliegenden Studie aufbauen können: Der wichtigste Baustein dazu sind Bildgebungsstudien, da diese das Fundament für die Entwicklung und Testung eines bildgebenden Biomarkers bilden. Bei diesen Studien sollte versucht werden, die Bildakquisition weiter zu vereinheitlichen, um den Einfluss von *scanner effects* zu minimieren. Zudem sollte die Gruppe aus Probanden im Hinblick auf mögliche Confounder für die Klassifizierung möglichst heterogen sein, um im späteren Modell *overfitting* zu vermeiden. Vor allem der T1-Classifier bedarf einer Überarbeitung, da seine Vorhersagegenauigkeit jener der Verfahren der anderen Autoren tendenziell unterlegen war. Mögliche Ansätze für eine Verbesserung sind u. a. eine andere Vorverarbeitung der Aufnahmen oder die Verwendung von *transfer learning*. Bei Anwendung eines prospektiven Studiendesigns könnte zudem ein Schwerpunkt auf Bereiche wie Früherkennung, Überwachung und Prognose des IPS oder gegebenenfalls verwandter Erkrankungen gelegt werden. Ein weiteres Thema für zukünftige Studien ist die Überprüfung der Classifier im Hinblick auf ihre Fähigkeit, zwischen Patienten mit IPS, Patienten mit Parkinsonismus ohne IPS oder solchen mit anderen Erkrankungen des zentralen Nervensystems zu differenzieren.

Möglicherweise tragen die Ergebnisse dieser Studie dazu bei, dass in zukünftigen Untersuchungen, unabhängig von IPS, die aus den rs-fMRT-Aufnahmen extrahierten Zeitserien vermehrt von den jeweiligen Autoren berücksichtigt werden. Dies ist wünschenswert, da diese zeitliche Komponente der Aufnahmen im Gegensatz zu den *spatial maps* bisher kaum Beachtung erhalten hat. Eine weitere sinnvolle Entwicklung bei medizinischen *machine learning*-Modellen ist die Verwendung von Alter und Geschlecht der Probanden als *features*. Diese Parameter sind einfach zu erheben und erlauben es dem Modell, alters- und geschlechterspezifische Eigenarten von Krankheiten zu berücksichtigen, wodurch sich schließlich die Vorhersagegenauigkeit steigern lässt.

6 Anhang

6.1 Literatur- und Quellenverzeichnis

Literaturverzeichnis

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Zugriff auf `https://www.tensorflow.org/` (Software available from tensorflow.org)

Abós, A., Baggio, H. C., Segura, B., García-Díaz, A. I., Compta, Y., Martí, M. J., ... Junqué, C. (2017). Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning. *Scientific reports*, *7* (1), 1–13.

Adler, C. H., Beach, T. G., Hentz, J. G., Shill, H. A., Caviness, J. N., Driver-Dunckley, E., ... Belden, C. M. (2014). Low clinical diagnostic accuracy of early vs advanced Parkinson disease: Clinicopathologic study. *Neurology*, *83* (5), 406–412.

Agniel, D., Kohane, I. S. und Weber, G. M. (2018). Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *Bmj*, *361*, k1479.

Alahmari, S. S., Goldgof, D. B., Mouton, P. R. und Hall, L. O. (2020). Challenges for the repeatability of deep learning models. *IEEE Access*, *8*, 211860–211868.

Alexander, G. E. und Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in neurosciences*, *13* (7), 266–271.

Andersson, J. L., Jenkinson, M. und Smith, S. (2007). Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2. *FMRIB Analysis Group of the University of Oxford*, *2* (1), e21.

Andersson, J. L. und Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage*, *125*, 1063–1078.

Andrews-Hanna, J. R., Smallwood, J. und Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, *1316* (1), 29.

Ardekani, B. A., Bachman, A. H., Strother, S. C., Fujibayashi, Y. und Yonekura, Y. (2004). Impact of inter-subject image registration on group analysis of fMRI data. In *International congress series* (Bd. 1265, S. 49–59).

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., ... Zeidman, P. (2014). SPM12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, 2464.

Ashburner, J. und Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage*, *11* (6), 805–821.

Ashburner, J. und Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26* (3), 839–851.

Badea, L., Onu, M., Wu, T., Roceanu, A. und Bajenaru, O. (2017). Exploring the reproducibility of functional connectivity alterations in Parkinson's disease. *PLoS One*, *12* (11), e0188196.

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., ... Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, *2* (1), 1–10.

Bastos, A. M. und Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, *9*, 175.

Baudrexel, S., Witte, T., Seifried, C., von Wegner, F., Beissner, F., Klein, J. C., ... Hilker, R. (2011). Resting state fMRI reveals increased subthalamic nucleus–motor cortex connectivity in Parkinson's disease. *Neuroimage*, *55* (4), 1728–1738.

Bayer, J. M., Thompson, P. M., Ching, C. R., Liu, M., Chen, A., Panzenhagen, A. C., ... Sämann, P. G. (2022). Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. *Frontiers in Neurology*, *13*, 923988.

Beavan, M., McNeill, A., Proukakis, C., Hughes, D. A., Mehta, A. und Schapira, A. H. (2015). Evolution of prodromal clinical markers of Parkinson disease in a GBA mutation–positive cohort. *JAMA neurology*, *72* (2), 201–208.

Beckmann, C. F., Mackay, C. E., Filippini, N. und Smith, S. M. (2009). Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *Neuroimage*, *47* (Suppl 1), S148.

Bell, A. J. und Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7* (6), 1129–1159.

Bhan, A., Kapoor, S., Gulati, M. und Goyal, A. (2021). Early Diagnosis of Parkinson's Disease in brain MRI using Deep Learning Algorithm. In *2021 third international conference on intelligent communication technologies and virtual mobile networks (icicv)* (S. 1467–1470).

Biswal, B., Zerrin Yetkin, F., Haughton, V. M. und Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic resonance in medicine*, *34* (4), 537–541.

Blum, A. L. und Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, *97* (1-2), 245–271.

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Lüders, E. und Rottenberg, D. (2004). Quantitative comparison of four brain extraction algorithms. *NeuroImage*, *22* (3), 1255–1261.

Brett, M., Markiewicz, C. J., Hanke, M., Côté, M.-A., Cipollini, B., McCarthy, P., ... freec84 (2020, November). *nipy/nibabel: 3.2.1.* Zenodo. Zugriff auf `https://doi.org/10.5281/zenodo.4295521` doi: 10.5281/zenodo.4295521

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78* (1), 1–3.

Brodersen, K. H., Ong, C. S., Stephan, K. E. und Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (S. 3121–3124).

Buch, V. H., Ahmed, I. und Maruthappu, M. (2018). Artificial intelligence in medicine: Current trends and future possibilities. *British Journal of General Practice*, *68* (668), 143–144.

Cai, L., Gao, J. und Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, *8* (11).

Carvalho, D. V., Pereira, E. M. und Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8* (8), 832.

Caspers, J., Mathys, C., Hoffstaedter, F., Südmeyer, M., Cieslik, E. C., Rubbert, C., ... Eickhoff, S. B. (2017). Differential functional connectivity alterations of two subdivisions within the right dlPFC in Parkinson's disease. *Frontiers in human neuroscience*, *11*, 288.

Caspers, J., Rubbert, C., Eickhoff, S. B., Hoffstaedter, F., Südmeyer, M., Hartmann, C. J., ... Mathys, C. (2021). Within-and across-network alterations of the sensorimotor network in Parkinson's disease. *Neuroradiology*, *63* (12), 2073–2085.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538* (7623), 20.

Chakraborty, S., Aich, S. und Kim, H.-C. (2020). Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network. *Diagnostics*, *10* (6), 402.

Chandna, P., Miron, M., Janer, J. und Gómez, E. (2017). Monoaural audio source separation using deep convolutional neural networks. In *International conference on latent variable analysis and signal separation* (S. 258–266).

Chen, Z. und Calhoun, V. (2018). Effect of spatial smoothing on task fMRI ICA and functional connectivity. *Frontiers in neuroscience*, *12*, 15.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15* (141), 20170387.

Chollet, F. (2015). *Keras.* GitHub. Zugriff auf `https://github.com/fchollet/keras`

Chollet, F. (2018). *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek.* MITP-Verlags GmbH & Co. KG.

Ciregan, D., Meier, U. und Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 ieee conference on computer vision and pattern recognition* (S. 3642–3649).

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A. und Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *Journal of Neuroscience*, *32* (26), 8988–8999.

Cordes, D., Haughton, V. M., Arfanakis, K., Carew, J. D., Turski, P. A., Moritz, C. H., ... Meyerand, M. E. (2001). Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data. *American Journal of Neuroradiology*, *22* (7), 1326–1333.

Cortes, C. und Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20* (3), 273–297.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., ... Clozel, T. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, *25* (10), 1519–1525.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20* (2), 215–232.

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'ardenne, K., ... Haxby, J. (2009). Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences*, *106* (26), 10415–10422.

De Graaf, M. A., Jager, K. J., Zoccali, C. und Dekker, F. W. (2011). Matching, an appealing method to avoid confounding? *Nephron Clinical Practice*, *118* (4), c315–c318.

Delenclos, M., Jones, D. R., McLean, P. J. und Uitti, R. J. (2016). Biomarkers in Parkinson's disease: Advances and strategies. *Parkinsonism & related disorders*, *22*, S106–S110.

Dorsey, E., Sherer, T., Okun, M. S. und Bloem, B. R. (2018). The emerging evidence of the Parkinson pandemic. *Journal of Parkinson's disease*, *8* (s1), S3–S8.

Eickhoff, S. B. und Grefkes, C. (2011). Approaches for the integrated analysis of structure, function and connectivity of the human brain. *Clinical EEG and neuroscience*, *42* (2), 107–121.

Eskofier, B. M., Lee, S. I., Daneault, J.-F., Golabchi, F. N., Ferreira-Carvalho, G., Vergara-Diaz, G., ... Bonato, P. (2016). Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment. In *2016 38th annual international conference of the ieee engineering in medicine and biology society (embc)* (S. 655–658).

Esmaeilzadeh, S., Yang, Y. und Adeli, E. (2018). End-to-end Parkinson disease diagnosis using brain MR-images by 3D-CNN. *arXiv preprint arXiv:1806.05233*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. und Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542* (7639), 115–118.

Evans, A. C., Collins, D. L., Mills, S., Brown, E. D., Kelly, R. L. und Peters, T. M. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In *1993 ieee conference record nuclear science symposium and medical imaging conference* (S. 1813–1817).

Fein, G., Di Sclafani, V., Taylor, C., Moon, K., Barakos, J., Tran, H., ... Shumway, R. (2004). Controlling for premorbid brain size in imaging studies: T1-derived cranium scaling factor vs. T2-derived intracranial vault volume. *Psychiatry Research: Neuroimaging*, *131* (2), 169–176.

Ferrari, E., Bosco, P., Calderoni, S., Oliva, P., Palumbo, L., Spera, G., ... Retico, A. (2020). Dealing with confounders and outliers in classification medical studies: The Autism Spectrum Disorders case study. *Artificial Intelligence in Medicine*, 101926.

Filippi, M., Sarasso, E., Piramide, N., Stojkovic, T., Stankovic, I., Basaia, S., ... Agosta, F. (2020). Progressive brain atrophy and clinical evolution in Parkinson's disease. *NeuroImage: Clinical*, *28*, 102374.

Forster, E. und Lewy, F. (1912). Paralysis agitans. *Pathologische Anatomie. Handbuch der Neurologie*, *20*, 920–933.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., ... McGrath, P. J. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, *167*, 104–120.

Frederick, J. und Meijer, B. G. (2014). Brain MRI in Parkinson's disease. *Frontiers in bioscience*, *6*, 360–369.

Garg, S. und McKeown, M. J. (2019). Functional Data and Long Short-Term Memory Networks for Diagnosis of Parkinson's Disease. In *International workshop on machine learning in medical imaging* (S. 655–663).

Gaser, C. und Dahnke, R. (2016). CAT-a computational anatomy toolbox for the analysis of structural MRI data. *HBM*, *2016*, 336–348.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Goodfellow, I., Bengio, Y. und Courville, A. (2016). *Deep Learning*. MIT press.

Gorgolewski, K. J., Esteban, O., Burns, C., Ziegler, E., Pinsard, B., Madison, C., ... Ghosh, S. (2016, April). *Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.12.0-rc1.* Zugriff auf `http://dx.doi.org/10.5281/zenodo.50186` doi: 10.5281/zenodo.50186

Griffanti, L., Rolinski, M., Szewczyk-Krolikowski, K., Menke, R. A., Filippini, N., Zamboni, G., ... Mackay, C. E. (2016). Challenges in the reproducibility of clinical studies with resting state fMRI: An example in early Parkinson's disease. *Neuroimage*, *124*, 704–713.

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L. und Yang, Y. (2018). Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*.

Haller, S., Badoud, S., Nguyen, D., Garibotto, V., Lovblad, K. und Burkhard, P. (2012). Individual detection of patients with Parkinson disease using support vector machine analysis of diffusion tensor imaging data: Initial results. *American Journal of Neuroradiology*, *33* (11), 2123–2128.

Hänggi, J., Fövenyi, L., Liem, F., Meyer, M. und Jäncke, L. (2014). The hypothesis of neuronal interconnectivity as a function of brain size—a general organization principle of the human connectome. *Frontiers in human neuroscience*, *8*, 915.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585* (7825), 357–362. Zugriff auf `https://doi.org/10.1038/s41586-020-2649-2` doi: 10.1038/s41586-020-2649-2

Helmich, R. C., Derikx, L. C., Bakker, M., Scheeringa, R., Bloem, B. R. und Toni, I. (2010). Spatial remapping of cortico-striatal connectivity in Parkinson's disease. *Cerebral cortex*, *20* (5), 1175–1186.

Hochreiter, S. und Schmidhuber, J. (1997). LSTM can solve hard long time lag problems. In *Advances in neural information processing systems* (S. 473–479).

Hoehn, M. M. und Yahr, M. D. (1967). Parkinsonism: Onset, progression, and mortality. *Neurology*, *17* (5), 427–442.

Hopfinger, J. B., Büchel, C., Holmes, A. P. und Friston, K. J. (2000). A study of analysis parameters that influence the sensitivity of event-related fMRI analyses. *Neuroimage*, *11* (4), 326–333.

Horwitz, B. (2003). The elusive concept of brain connectivity. *Neuroimage*, *19* (2), 466–470.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. und Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18* (8), 500–510.

Hoyer, A. und Zapf, A. (2021). Studies for the Evaluation of Diagnostic Tests: Part 28 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, *118* (33-34), 555.

Huang, T.-W., Chen, H.-T., Fujimoto, R., Ito, K., Wu, K., Sato, K., ... Aoki, T. (2017). Age estimation from brain MRI images using deep learning. In *2017 ieee 14th international symposium on biomedical imaging (isbi 2017)* (S. 849–852).

Huettel, S. A., Singerman, J. D. und McCarthy, G. (2001). The effects of aging upon the hemodynamic response measured by functional MRI. *Neuroimage*, *13* (1), 161–175.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9* (3), 90–95. doi: 10.1109/MCSE.2007.55

Hustad, E., Skogholt, A. H., Hveem, K. und Aasly, J. O. (2018). The accuracy of the clinical diagnosis of Parkinson disease. The HUNT study. *Journal of Neurology*, *265*, 2120–2124.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, *10* (3), 626–634.

Ibarretxe-Bilbao, N., Junque, C., Marti, M. J. und Tolosa, E. (2011). Brain structural MRI correlates of cognitive dysfunctions in Parkinson's disease. *Journal of the neurological sciences*, *310* (1-2), 70–74.

Iheagwam, F. N. und Etefia, S. I. (2019). Recent Advances on the Management of Parkinson's Disease: A Review.

Iqbal, J. D. und Vinay, R. (2022). Are we ready for Artificial Intelligence in Medicine? *Swiss Medical Weekly* (19).

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., ... die Alzheimer's Disease Neuroimaging Initiative (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27* (4), 685–691.

Jankovic, J., Rajput, A. H., McDermott, M. P., Perl, D. P. und die Parkinson Study Group. (2000). The evolution of diagnosis in early Parkinson disease. *Archives of neurology*, *57* (3), 369–372.

Jenkinson, M., Bannister, P., Brady, M. und Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17* (2), 825–841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. und Smith, S. M. (2012). FSL. *Neuroimage*, *62* (2), 782–790.

Jenkinson, M. und Chappell, M. (2018). *Introduction to Neuroimaging Analysis*. Oxford University Press.

Jenkinson, M., Pechaud, M. und Smith, S. (2005). BET2: MR-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping* (Bd. 17, S. 167).

Jenkinson, M. und Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, *5* (2), 143–156.

Jiang, X., Osl, M., Kim, J. und Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, *19* (2), 263–274.

Jnawali, K., Arbabshirani, M. R., Rao, N. und Patel, A. A. (2018). Deep 3D convolution neural network for CT brain hemorrhage classification. In *Medical imaging 2018: Computer-aided diagnosis* (Bd. 10575, S. 307–313).

Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivières, S., Grigis, A., ... Whelan, R. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*, *199*, 351–365.

Jovicich, J., Minati, L., Marizzoni, M., Marchitelli, R., Sala-Llonch, R., Bartrés-Faz, D., ... Frisoni, G. B. (2016). Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: A multicentric resting-state fMRI study. *Neuroimage*, *124*, 442–454.

Kanda, T. und Uchida, S.-i. (2014). Clinical/pharmacological aspect of adenosine A2A receptor antagonist for dyskinesia. In *International review of neurobiology* (Bd. 119, S. 127–150). Elsevier.

Kelly, B. S., Judge, C., Bollard, S. M., Clifford, S. M., Healy, G. M., Aziz, A., ... Killeen, R. P. (2022). Radiology artificial intelligence: A systematic review and evaluation of methods (RAISE). *European radiology*, 1–10.

Kingma, D. P. und Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiviniemi, V., Kantola, J.-H., Jauhiainen, J., Hyvärinen, A. und Tervonen, O. (2003). Independent component analysis of nondeterministic fMRI signal sources. *Neuroimage*, *19* (2), 253–260.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., ... Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, *46* (3), 786–802.

Kowal, S. L., Dall, T. M., Chakrabarti, R., Storm, M. V. und Jain, A. (2013). The current and projected economic burden of Parkinson's disease in the United States. *Movement Disorders*, *28* (3), 311–318.

Krizhevsky, A., Sutskever, I. und Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kruggel, F., Turner, J., Muftuler, L. T. und die Alzheimer's Disease Neuroimaging Initiative. (2010). Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. *Neuroimage*, *49* (3), 2123–2133.

Lakhani, P. und Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, *284* (2), 574–582.

Lawrence, J., Malmsten, J., Rybka, A., Sabol, D. A. und Triplin, K. (2017). Comparing TensorFlow deep learning performance using CPUs, GPUs, local PCs and cloud.

Le Bihan, D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nature reviews neuroscience*, *4* (6), 469–480.

Lebovitz, S., Levina, N. und Lifshitz-Assaf, H. (2021). Is AI Ground Truth Really 'True'? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What (May 4, 2021). Citation: Lebovitz, S., Levina, N., Lifshitz-Assaf, H*, 1501–1525.

Lee, A. und Gilbert, R. M. (2016). Epidemiology of Parkinson disease. *Neurologic clinics*, *34* (4), 955–965.

Lee, C., Jang, J., Lee, S., Kim, Y. S., Jo, H. J. und Kim, Y. (2020). Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. *Scientific reports*, *10* (1), 1–12.

Lei, D., Chen, X. und Zhao, J. (2018). Opening the black box of deep learning. *arXiv preprint arXiv:1805.08355*.

Leung, K. K., Barnes, J., Modat, M., Ridgway, G. R., Bartlett, J. W., Fox, N. C., ... die Alzheimer's Disease Neuroimaging Initiative (2011). Brain MAPS: An automated, accurate and robust brain extraction technique using a template library. *Neuroimage*, *55* (3), 1091–1108.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D. und Chen, M. (2014). Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (icarcv)* (S. 844–848).

Li, X., Morgan, P. S., Ashburner, J., Smith, J. und Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of neuroscience methods*, *264*, 47–56.

Li, Z. und Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, *40* (12), 2935–2947.

Lindquist, M. A., Loh, J. M. und Yue, Y. R. (2010). Adaptive spatial smoothing of fMRI images. *Statistics and its Interface*, *3* (1), 3–13.

Logothetis, N. K. und Wandell, B. A. (2004). Interpreting the BOLD signal. *Annu. Rev. Physiol.*, *66*, 735–769.

Lowekamp, B. C., Chen, D. T., Ibáñez, L. und Blezek, D. (2013). The design of SimpleITK. *Frontiers in neuroinformatics*, *7*, 45.

Maas, L. C. und Renshaw, P. F. (1999). Post-registration spatial filtering to reduce noise in functional MRI data sets-A general linear approach. *Magnetic Resonance Imaging*, *9* (17), 1371–1382.

MacDonald, M. E. und Pike, G. B. (2021). MRI of healthy brain aging: A review. *NMR in Biomedicine*, *34* (9), e4564.

Malek, N., Swallow, D., Grosset, K., Anichtchik, O., Spillantini, M. und Grosset, D. (2014). Alpha-synuclein in peripheral tissues and body fluids as a biomarker for Parkinson's disease–a systematic review. *Acta Neurologica Scandinavica*, *130* (2), 59–72.

Marchini, J. L. und Ripley, B. D. (2000). A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, *12* (4), 366–380.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., ... die Parkinson Progression Marker Initiative (2011). The parkinson progression marker initiative (PPMI). *Progress in neurobiology*, *95* (4), 629–635.

Martuzzi, R., Ramani, R., Qiu, M., Shen, X., Papademetris, X. und Constable, R. T. (2011). A whole-brain voxel based measure of intrinsic connectivity contrast reveals local changes in tissue connectivity with anesthetic without a priori assumptions on thresholds or regions of interest. *Neuroimage*, *58* (4), 1044–1050.

Mathys, C., Caspers, J., Langner, R., Suedmeyer, M., Grefkes, C., Reetz, K., ... Eickhoff, S. B. (2016). Functional Connectivity Differences of the Subthalamic Nucleus Related to Parkinson's Disease. *Human brain mapping*, *37* (3), 1235–1253.

McCarthy, P. (2020, Juli). *FSLeyes.* Zenodo. Zugriff auf `https://doi.org/10.5281/zenodo.3937147` doi: 10.5281/zenodo.3937147

McKeown, M. J., Hansen, L. K. und Sejnowsk, T. J. (2003). Independent component analysis of functional MRI: What is signal and what is noise? *Current opinion in neurobiology*, *13* (5), 620–629.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J. und Sejnowski, T. J. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human brain mapping*, *6* (3), 160–188.

McNeill, A., Wu, R.-M., Tzen, K.-Y., Aguiar, P. C., Arbelo, J. M., Barone, P., ... Schapira, A. H. V. (2013). Dopaminergic neuronal imaging in genetic Parkinson's disease: Insights into pathogenesis. *PloS one*, *8* (7), e69190.

Miall, R. C. und Robertson, E. M. (2006). Functional imaging: Is the resting brain resting? *Current Biology*, *16* (23), R998–R1000.

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., ... Krupa, P. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic resonance imaging*, *26* (4), 490–503.

Milam, M. und Koo, C. (2023). The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. *Clinical Radiology*, *78* (2), 115–122.

Ngan, S.-C., LaConte, S. M. und Hu, X. (2000). Temporal filtering of event-related fMRI data using cross-validation. *NeuroImage*, *11* (6), 797–804.

Nickerson, L. D., Smith, S. M., Öngür, D. und Beckmann, C. F. (2017). Using dual regression to investigate network shape and amplitude in functional connectivity analyses. *Frontiers in neuroscience*, *11*, 115.

Nisha, R. M. und Sharma, L. (2015). Comparative analysis of canny and prewitt edge detection techniques used in image processing. *International Journal of Engineering Trends and Technology*, *28* (1), 48–53.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G. und Tran, D. (2019). Measuring Calibration in Deep Learning. In *Cvpr workshops* (Bd. 2).

Noble, S., Scheinost, D. und Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, *203*, 116157.

Nooner, K. B., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., ... Milham, M. P. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, *6*, 152.

Nunes, A., Schnack, H. G., Ching, C. R., Agartz, I., Akudjedu, T. N., Alda, M., ... die ENIGMA Bipolar Disorders Working Group (2020). Using structural MRI to identify bipolar disorders–13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Molecular psychiatry*, *25* (9), 2130–2143.

Nyúl, L. G. und Udupa, J. K. (1999). On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *42* (6), 1072–1081.

Ogawa, T., Fujii, S., Kuya, K., Kitao, S.-i., Shinohara, Y., Ishibashi, M. und Tanabe, Y. (2018). Role of neuroimaging on differentiation of Parkinson's disease and its related diseases. *Yonago acta medica*, *61* (3), 145–155.

Oh, S. L., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M. und Acharya, U. R. (2018). A deep learning approach for Parkinson's disease diagnosis from EEG signals. *Neural Computing and Applications*, 1–7.

Palma, J.-A., Norcliffe-Kaufmann, L. und Kaufmann, H. (2018). Diagnosis of multiple system atrophy. *Autonomic Neuroscience*, *211*, 15–25.

Park, J. G. und Lee, C. (2009). Skull stripping based on region growing for magnetic resonance brain images. *NeuroImage*, *47* (4), 1394–1407.

Pawlowski, N. und Glocker, B. (2019). Is Texture Predictive for Age and Sex in Brain MRI? *arXiv preprint arXiv:1907.10961*.

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

Pereira, C. R., Weber, S. A., Hook, C., Rosa, G. H. und Papa, J. P. (2016). Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics. In *2016 29th sibgrapi conference on graphics, patterns and images (sibgrapi)* (S. 340–346).

Peters, A. und Morrison, J. H. (2012). *Cerebral cortex: Neurodegenerative and age-related changes in structure and function of cerebral cortex* (Bd. 14). Springer Science & Business Media.

Pohjalainen, T., Rinne, J. O., Någren, K., Syvälahti, E. und Hietala, J. (1998). Sex differences in the striatal dopamine D2 receptor binding characteristics in vivo. *American Journal of Psychiatry*, *155* (6), 768–773.

Polat, H. und Danaei Mehr, H. (2019). Classification of pulmonary CT images by using hybrid 3D-deep convolutional neural network architecture. *Applied Sciences*, *9* (5), 940.

Politis, M., Pagano, G. und Niccolini, F. (2017). Imaging in Parkinson's disease. In *International review of neurobiology* (Bd. 132, S. 233–274). Elsevier.

Pool, E.-M., Rehme, A. K., Eickhoff, S. B., Fink, G. R. und Grefkes, C. (2015). Functional resting-state connectivity of the human motor network: Differences between right-and left-handers. *NeuroImage*, *109*, 298–306.

Pooley, R. A. (2005). Fundamental physics of MR imaging. *Radiographics*, *25* (4), 1087–1099.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. und Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, *59* (3), 2142–2154.

Power, J. D., Schlaggar, B. L. und Petersen, S. E. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage*, *105*, 536–551.

Prashanth, R., Roy, S. D., Mandal, P. K. und Ghosh, S. (2016). High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. *International journal of medical informatics*, *90*, 13–21.

Prasuhn, J., Heldmann, M., Münte, T. F. und Brüggemann, N. (2020). A machine learning-based classification approach on Parkinson's disease diffusion tensor imaging datasets. *Neurological Research and Practice*, *2* (1), 1–5.

Quattrone, A., Morelli, M., Nigro, S., Quattrone, A., Vescio, B., Arabia, G., ... Novellino, F. (2018). A new MR imaging index for differentiation of progressive supranuclear palsy-parkinsonism from Parkinson's disease. *Parkinsonism & related disorders*, *54*, 3–8.

Rajapakse, J. C., Giedd, J. N. und Rapoport, J. L. (1997). Statistical approach to segmentation of single-channel cerebral MR images. *IEEE transactions on medical imaging*, *16* (2), 176–186.

Ramírez, V. M., Forbes, F., Coupé, P. und Dojat, M. (2019). No structural Brain differences in'de novo'Parkinsonian patients..

Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K. und Wu, Z. (2019). *Deep Learning for the Life Sciences*. O'Reilly Media.

Raudys, S. J. und Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, *13* (3), 252–264.

Ravina, B., Marek, K., Eberly, S., Oakes, D., Kurlan, R., Ascherio, A., ... Galpern, W. R. (2012). Dopamine transporter imaging is associated with long-term outcomes in Parkinson's disease. *Movement Disorders*, *27* (11), 1392–1397.

Reback, J., McKinney, W., jbrockmendel, den Bossche, J. V., Augspurger, T., Cloud, P., ... h vetinari (2021, März). *pandas-dev/pandas: Pandas 1.2.3*. Zenodo. Zugriff auf `https://doi.org/10.5281/zenodo.4572994` doi: 10.5281/zenodo.4572994

Reese, J., Winter, Y., Balzer-Geldsetzer, M., Bötzel, K., Eggert, K., Oertel, W., ... von Campenhausen, S. (2011). Morbus Parkinson: Krankheitskosten einer ambulanten Patientenkohorte. *Das Gesundheitswesen*, *73* (01), 22–29.

Ribeiro, M. T., Singh, S. und Guestrin, C. (2016). 'Why should i trust you?' Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (S. 1135–1144).

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A. und Logroscino, G. (2016). Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology*, *86* (6), 566–576.

Rubbert, C., Mathys, C., Jockwitz, C., Hartmann, C. J., Eickhoff, S. B., Hoffstaedter, F., ... Caspers, J. (2019). Machine-learning identifies Parkinson's disease patients based on resting-state between-network functional connectivity. *The British journal of radiology*, *92* (1101), 20180886.

Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K. und Kaski, K. (2019). Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific reports*, *9* (1), 1–11.

Saponaro, S., Giuliano, A., Bellotti, R., Lombardi, A., Tangaro, S., Oliva, P., ... Retico, A. (2022). Multi-site harmonization of MRI data uncovers machine-learning discrimination capability in barely separable populations: An example from the ABIDE dataset. *NeuroImage: Clinical*, 103082.

Sarasso, E., Agosta, F., Piramide, N. und Filippi, M. (2021). Progression of grey and white matter brain damage in Parkinson's disease: A critical review of structural MRI literature. *Journal of neurology*, *268* (9), 3144–3179.

Scarpazza, C. und De Simone, M. S. (2016). Voxel-based morphometry: Current perspectives. *Neuroscience and Neuroeconomics*, *5*, 19–35.

Schrag, A., Horsfall, L., Walters, K., Noyce, A. und Petersen, I. (2015). Prediagnostic presentations of Parkinson's disease in primary care: A case-control study. *The Lancet Neurology*, *14* (1), 57–64.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. und Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the ieee international conference on computer vision* (S. 618–626).

Shaheen, F., Verma, B. und Asafuddoula, M. (2016). Impact of automatic feature extraction in deep learning architecture. In *2016 international conference on digital image computing: Techniques and applications (dicta)* (S. 1–8).

Sharman, M., Valabregue, R., Perlbarg, V., Marrakchi-Kacem, L., Vidailhet, M., Benali, H., ... Lehéricy, S. (2013). Parkinson's disease patients show reduced cortical-subcortical sensorimotor connectivity. *Movement Disorders*, *28* (4), 447–454.

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R. und Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, *9* (1), 1–12.

Shi, D., Zhang, H., Wang, G., Wang, S., Yao, X., Li, Y., ... Ren, K. (2022). Machine Learning for Detecting Parkinson's Disease by Resting-State Functional Magnetic Resonance Imaging: A Multicenter Radiomics Analysis. *Frontiers in aging neuroscience*, *14*, 806828.

Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P. K. und Ingalhalikar, M. (2019). Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. *NeuroImage: Clinical*, *22*, 101748.

Sivaranjini, S. und Sujatha, C. (2020). Deep learning based diagnosis of Parkinson's disease using convolutional neural network. *Multimedia tools and applications*, *79* (21), 15467–15479.

Smith, A. M., Lewis, B. K., Ruttimann, U. E., Frank, Q. Y., Sinnwell, T. M., Yang, Y., ... Frank, J. A. (1999). Investigation of low frequency drift in fMRI signal. *Neuroimage*, *9* (5), 526–533.

Smith, A. T., Singh, K. D. und Balsters, J. H. (2007). A comment on the severity of the effects of non-white noise in fMRI time-series. *NeuroImage*, *36* (2), 282–288.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, *17* (3), 143–155.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208–S219.

Spix, C. und Blettner, M. (2012). Screening: Part 19 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *109* (21), 385.

Strimbu, K. und Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5* (6), 463.

Sullivan, G. M. und Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education*, *4* (3), 279–282.

Sultana, F., Sufian, A. und Dutta, P. (2018). Advancements in image classification using convolutional neural network. In *2018 fourth international conference on research in computational intelligence and communication networks (icrcicn)* (S. 122–129).

Summerfield, C., Junqué, C., Tolosa, E., Salgado-Pineda, P., Gómez-Ansón, B., Martí, M. J., ... Mercader, J. (2005). Structural brain changes in Parkinson disease with dementia: A voxel-based morphometry study. *Archives of neurology*, *62* (2), 281–285.

Talai, A. S., Sedlacik, J., Boelmans, K. und Forkert, N. D. (2021). Utility of Multi-Modal MRI for Differentiating of Parkinson's Disease and Progressive Supranuclear Palsy Using Machine Learning. *Frontiers in Neurology*, 546.

Tavares, V., Prata, D. und Ferreira, H. A. (2020). Comparing SPM12 and CAT12 segmentation pipelines: A brain tissue volume-based age and Alzheimer's disease study. *Journal of Neuroscience Methods*, *334*, 108565.

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K. und Brink, J. (2018). Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology*, *15* (3), 504–508.

Thulborn, K. R., Waterton, J. C., Matthews, P. M. und Radda, G. K. (1982). Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *714* (2), 265–270.

Tian, Z.-y., Qian, L., Fang, L., Peng, X.-h., Zhu, X.-h., Wu, M., ... Shao, J. (2020). Frequency-specific changes of resting brain activity in Parkinson's disease: A machine learning approach. *Neuroscience*, *436*, 170–183.

Tolosa, E., Wenning, G. und Poewe, W. (2006). The diagnosis of Parkinson's disease. *The Lancet Neurology*, *5* (1), 75–86.

Ulla, M., Bonny, J. M., Ouchchane, L., Rieu, I., Claise, B. und Durif, F. (2013). Is R2* a new MRI biomarker for the progression of Parkinson's disease? A longitudinal follow-up. *PLoS One*, *8* (3).

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E. und Ugurbil, K. (2013). The WU-Minn human connectome project: An overview. *Neuroimage*, *80*, 62–79.

Van Rossum, G. und Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Eskofier, B., Klucken, J. und Nöth, E. (2018). Multimodal assessment of Parkinson's disease: A deep learning approach. *IEEE journal of biomedical and health informatics*, *23* (4), 1618–1630.

Vieira, S., Pinaya, W. H. und Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58–75.

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

Wachinger, C., Becker, B. G., Rieckmann, A. und Pölsterl, S. (2019). Quantifying confounding bias in neuroimaging datasets with causal inference. In *International conference on medical image computing and computer-assisted intervention* (S. 484–492).

Wang, J., Zhu, H., Wang, S.-H. und Zhang, Y.-D. (2021). A review of deep learning on medical image analysis. *Mobile Networks and Applications*, *26* (1), 351–380.

Warfield, S., Robatino, A., Dengler, J., Jolesz, F. und Kikinis, R. (1999). Nonlinear registration and template driven segmentation. *Brain warping*, *4*, 67–84.

Waskom, M. L. (2021). seaborn: Statistical Data Visualization. *Journal of Open Source Software*, *6* (60), 3021. Zugriff auf `https://doi.org/10.21105/joss.03021` doi: 10.21105/joss.03021

Welvaert, M. und Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. *PloS one*, *8* (11), e77089.

Wenckebach, T. H. (2005). *Volumetrische Registrierung zur medizinischen Bildanalyse* (Unveröffentlichte Dissertation). Master's thesis, Institut für Informatik, Humboldt-Universität zu Berlin.

Whitcher, B., Schmid, V. J. und Thornton, A. (2011). Working with the DICOM and NIfTI Data Standards in R. *Journal of Statistical Software* (6).

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., ... Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, *45* (1), S173–S186.

Woolrich, M. W., Ripley, B. D., Brady, M. und Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, *14* (6), 1370–1386.

WU-Minn HCP. (2017). 1200 subjects data release reference manual. *URL https://www.humanconnectome. org*.

Xu, Y., Kong, Q., Wang, W. und Plumbley, M. D. (2018). Large-scale weakly supervised audio classification using gated convolutional neural network. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (S. 121–125).

Yan, W., Calhoun, V., Song, M., Cui, Y., Yan, H., Liu, S., ... Sui, J. (2019). Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data. *EBioMedicine*, *47*, 543–552.

Yari, Y., Nguyen, T. V. und Nguyen, H. T. (2020). Deep learning applied for histological diagnosis of breast cancer. *IEEE Access*, *8*, 162432–162448.

Yasaka, K., Kamagata, K., Ogawa, T., Hatano, T., Takeshige-Amano, H., Ogaki, K., ... Abe, O. (2021). Parkinson's disease: Deep learning with a parameter-weighted structural connectome matrix for diagnosis and neural circuit disorder investigation. *Neuroradiology*, *63* (9), 1451–1462.

Zach, H., Walter, U., Liepelt-Scarfone, I. und Maetzler, W. (2017). Diagnostik des klinischen und prodromalen idiopathischen Parkinson-Syndroms. *Der Nervenarzt*, *88* (4), 356–364.

Zadrozny, B. und Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (S. 694–699).

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J. und Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*.

Zhao, H., Tsai, C.-C., Zhou, M., Liu, Y., Chen, Y.-L., Huang, F., ... Wang, J.-J. (2022). Deep learning based diagnosis of Parkinson's Disease using diffusion magnetic resonance imaging. *Brain Imaging and Behavior*, 1–12.

Zhao, Q., Adeli, E. und Pohl, K. M. (2020). Training confounder-free deep learning models for medical applications. *Nature communications*, *11* (1), 6010.

7 Danksagung

Mein größter Dank gebührt meinem Doktorvater Prof. Dr. med. Dipl.-Inform. Julian Caspers, der meine Studienzeit wesentlich beeinflusst hat, indem er mein Interesse an neuronale Bildgebung geweckt und stets gefördert hat. Neben seiner fachlichen Kompetenz unterstützte er mich auch emotional in schwierigen Phasen des Projekts und ermutigte mich dazu, so lange neue Ideen auszuprobieren und weiter auszubauen, bis schließlich die vorliegenden Ergebnisse entstanden sind.

Des Weiteren danke ich meinem Co-Betreuer Prof. Dr. med. Simon B. Eickhoff, der sowohl durch kritisches Hinterfragen meiner Vorgehensweise als auch durch das Vorschlagen neuer, kreativer Ideen die Entwicklung der vorliegenden Studie wesentlich vorangetrieben und positiv beeinflusst hat.

Meinem Betreuer PD Dr. med. Christian Rubbert danke ich zudem für seine Unterstützung bei allen möglichen Herausforderungen während meiner Forschung und darüber hinaus. Insbesondere bei der Verwendung verschiedener Datenbanken und der Bewältigung zahlreicher technischer Hürden konnte ich mich immer auf ihn verlassen.

Ein besonderer Dank gilt meinem Bruder Stefan Boschenriedter, der mich schon lange vor Beginn der Dissertation umfangreich in Keras sowie weitere Python-Bibliotheken eingearbeitet und bei technischen Problemen jederzeit unterstützt hat. Ohne ihn hätte es die vorliegende Studie in dieser Qualität vermutlich nicht gegeben.

Die rechnerische Infrastruktur für die Vorverarbeitung der strukturellen T1-gewichteten Aufnahmen mit CAT12 wurden vom Zentrum für Informations- und Medientechnologie der Heinrich-Heine-Universität Düsseldorf bereitgestellt. Insbesondere Dipl.-Inform. Christian Siebert danke ich dabei für die sehr freundliche und kompetente Unterstützung bei der Verwendung des Hochleistungs-Rechenclusters HILBERT.

Die für die Erstellung der vorliegenden Studie verwendeten Datensätze stammen u. a. aus der Datenbank der Parkinson's Progression Markers Initiative Datenbank (`ppmi-info.org/access-data-specimens/download-data`). Aktuelle Informationen über diese Studie sind unter `ppmi-info.org` abrufbar. Die PPMI ist eine öffentlich-private Partnerschaft und wird von der Michael J. Fox Foundation for Parkinson's Research mit den Partnern 4D Pharma, Abbvie, Acurex Therapeutics, Allergan, Amathus Therapeutics, ASAP, Avid Radiopharmaceuticals, Bial Biotech, Biogen, BioLegend, Bristol-Myers Squibb, Calico, Celgene, Dacapo Brain Science, Denali, The Edmond J. Safra Foundaiton, GE Healthcare, Genentech, GlaxoSmithKline, Golub Capital, Handl Therapeutics, Insitro, Janssen Neuroscience, Lilly, Lundbeck, Merck, Meso Scale Discovery, Neurocrine Biosciences, Pfizer, Piramal, Prevail, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UCB, Verily und Voyager Therapeutics finanziert. Weitere Finanzierungspartner sind Industrieunternehmen, gemeinnützige Organisationen und Privatpersonen (`ppmi-info.org/about-ppmi/who-we-are/study-sponsors`).

Es wurden außerdem Datensätze aus der Datenbank der Alzheimer's Disease Neuroimaging Initiative (ADNI, `adni.loni.usc.edu`) verwendet. Die ADNI wurde im Jahr 2003 als öffentlich-private Partnerschaft unter der Leitung von Michael W. Weiner gegründet.

Das Hauptziel von ADNI bestand darin, zu prüfen, ob MRT-Aufnahmen, die Positronen-Emissions-Tomographie, andere biologische Marker sowie klinische und neuropsychologische Beurteilung kombiniert werden können, um das Fortschreiten von leichten kognitiven Beeinträchtigung und der frühen Alzheimer-Krankheit zu messen. Aktuelle Informationen sind auf `adni-info.org` zu finden. Die Datenerfassung und -weitergabe für dieses Projekt wurde von der Alzheimer's Disease Neuroimaging Initiative (ADNI, National Institutes of Health Grant U01 AG024904) und DOD ADNI (Department of Defense award number W81XWH-12-2-0012) finanziert. ADNI wird durch das National Institute on Aging, das National Institute of Biomedical Imaging and Bioengineering und durch großzügige Spenden der folgenden Organisationen finanziert: AbbVie, Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Araclon Biotech, BioClinica, Biogen, Bristol-Myers Squibb Company, CereSpir, Cogstate, Eisai Inc, Elan Pharmaceuticals, Eli Lilly and Company, EuroImmun, F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Fujirebio, GE Healthcare, IXICO Ltd., Janssen Alzheimer Immunotherapy Research & Development, Johnson & Johnson Pharmaceutical Research & Development LLC., Lumosity, Lundbeck, Merck & Co., Meso Scale Diagnostics, NeuroRx Research, Neurotrack Technologies, Novartis Pharmaceuticals Corporation, Pfizer Inc., Piramal Imaging, Servier, Takeda Pharmaceutical Company und Transition Therapeutics. Das Canadian Institutes of Health Research stellt Mittel zur Unterstützung der klinischen Zentren von ADNI in Kanada zur Verfügung. Beiträge aus dem privaten Sektor werden von der Foundation for the National Institutes of Health (`fnih.org`) zur Verfügung gestellt. Die Stipendiatenorganisation ist das Northern California Institute for Research and Education und die Studie wird vom Alzheimer's Therapeutic Research Institute an der University of Southern California koordiniert. Die ADNI-Daten werden vom Laboratory for Neuro Imaging an der University of Southern California bereitgestellt.