UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF

Klinik und Poliklinik für Neuroradiologische Diagnostik und Intervention

Direktor: Prof. Dr. med. Jens Fiehler

Integration von räumlichen Informationen in voxelbasierte Modelle zur Vorhersage des Gewebeschicksals beim ischämischen Schlaganfall

Dissertation

zur Erlangung des Doktorgrades Dr. rer. biol. hum.
an der Medizinischen Fakultät der Universität Hamburg

vorgelegt von:

Malte Grosser
aus Meppen

Hamburg 2022

Angenommen von der Medizinischen Fakultät am: **20.09.2022**

Veröffentlicht mit Genehmigung der Medizinischen Fakultät der Universität Hamburg

Prüfungsausschuss, der/die Vorsitzende: **Prof. Dr. Jens Fiehler**

Prüfungsausschuss, 2. Gutachter/in: **Prof. Dr. Karl Wegscheider**

Abstract

The advantages and risks of therapy is a case-by-case evaluation no later than 4.5 or 6 hours after the onset of an ischaemic stroke. The assessment whether a recanalizing therapy based on perfusion-diffusion mismatch is potentially the right treatment choice has proven to be less reliable. Therefore, machine learning methods are increasingly investigated to predict tissue outcome in ischaemic stroke. Although lesions are constrained by arterial territories and are thus not randomly distributed, spatial factors have so far been mostly ignored for prediction.

The aim of this thesis is to determine whether the location of a voxel and the previously known relative lesion frequency contribute to better predictions of tissue fate. For that purpose, two approaches using these spatial information for prediction were tested in 99 acute stroke patients from the I-KNOW study. To ensure comparability of voxel positions, the individual MRT data sets of all patients were mapped to the Montreal Neurological Institute Atlas which served as a reference. For local modelling, one model was trained for each position. In the event of insufficient data, a global model was used, which had been trained on voxels from all positions. A hybrid model was generated from the average of the global and local approaches to combine their complementary predictions. To integrate the spatial information into a global model, the a priori known lesion frequencies and voxel coordinates served as features. Logistic regression and random forest as well as XGBoost were used as algorithms.

The predictions of all models were evaluated in a leave-one-patient-out cross-validation using the area under the receiver operating characteristic curve and the Dice coefficient. The hybrid approach demonstrated significant improvements over the local and global approaches. The integration of spatial features resulted in statistically meaningful improvements, especially for tree-based models. The inclusion of spatial features resulted in a substantial optimization of predicting the tissue outcome because the models learned to differentiate the endangered arterial supply areas of the anterior territory from the rest of the tissue. The specific site of occlusion and the voxel locations described in this thesis contribute to a more individualized delineation of the tissue margins, which should be both integrated simultaneously in future studies.

Zusammenfassung

Beim ischämischen Schlaganfall müssen nach Ablauf von 4,5 bzw. 6 Stunden Vor- und Nachteile einer Therapie fallspezifisch bewertet werden. Eine Einschätzung des Potenzials einer rekanalisierenden Therapie auf Basis des Perfusions-Diffusions-Mismatches erweist sich häufig als optimierungsbedürftig. Deshalb werden zunehmend Methoden des maschinellen Lernens zur Vorhersage des Gewebeoutcomes beim ischämischen Schlaganfall erforscht. Obwohl Läsionen durch arterielle Territorien eingeschränkt und nicht zufällig verteilt sind, fließen räumliche Faktoren bislang nur selten in die Prädiktion ein.

Mit dieser Arbeit soll geklärt werden, ob der Einbezug der Lage eines Voxels und die lageabhängige a priori bekannte relative Läsionshäufigkeit zu einer verbesserten Vorhersage des Gewebeschicksals in voxelbasierten Modellen führt. Hierzu wurden zwei Ansätze anhand von 99 akuten Schlaganfallpatienten aus der I-KNOW-Studie erprobt, die die räumlichen Informationen berücksichtigen. Zur Vergleichbarkeit von Voxelpositionen fand eine Registrierung aller Bilddaten auf den Montreal Neurological Institute Gehirnatlas statt, der als Referenz diente. Für die lokale Modellierung wurde ein Modell pro Position trainiert. Bei unzureichender Datenmenge wurde an diesen Positionen ein globales Modell herangezogen, das auf Voxeln aller Positionen trainiert wurde. Die komplementären Vorhersagen dieser Ansätze wurden über den Durchschnitt in einem hybriden Modell vereint. Zur Integration der räumlichen Informationen in ein globales Modell dienten die a priori bekannten Läsionshäufigkeiten und Voxelkoordinaten als Features. Als Algorithmen kamen die logistische Regression, Random Forest sowie XGBoost zum Einsatz.

Ausgewertet wurden die Vorhersagen aller Modelle im Rahmen einer Leave-One-Patient-Out-Kreuzvalidierung in Bezug auf die Fläche unter der Receiver-Operating-Characteristic-Kurve und den Dice-Koeffizienten. Dabei zeigte der hybride Ansatz signifikante Verbesserungen gegenüber dem lokalen und globalen Ansatz. Die Integration von räumlichen Features in den globalen Ansatz führte vor allem bei tree-basierten Modellen zu statistisch relevanten Verbesserungen. Demnach trägt der Einbezug räumlicher Informationen zu einer deutlich optimierten Vorhersage des Gewebeschicksals von Patienten mit akutem Schlaganfall bei, denn die Modelle lernten die gefährdeten arteriellen Versorgungsgebiete des vorderen Stromgebiets vom Rest des Gewebes zu differenzieren. Zur individuelleren Abgrenzung des Gewebes trägt neben den in dieser Arbeit integrierten Voxelpositionen auch noch die Lage des Verschlusses bei, die es beide gleichzeitig in zukünftigen Studien zu integrieren gilt.

Inhaltsverzeichnis

Abbildungsverzeichnis V

Tabellenverzeichnis VI

Abkürzungsverzeichnis VII

1 Einleitung 1

1.1 Zielsetzung 2

1.2 Aufbau der Arbeit 3

2 Diagnostische Verfahren zur Intervention beim akuten ischämischen Schlaganfall 7

2.1 Grundlagen 7

2.1.1 Ätiologie 8

2.1.2 Epidemiologie 10

2.1.3 Risikofaktoren 12

2.1.4 Die arterielle Blutversorgung des Gehirns 13

2.2 Akute Symptomatik und Gewebeschäden 15

2.3 Rekanalisationstherapien zur Rettung der Penumbra 18

2.4 Bildgebung mittels Magnetresonanztomographie 20

2.4.1 Mechanismen der Magnetresonanztomographie 21

2.4.2 MRT-Sequenzen zur Auswahl der Therapieverfahren 27

2.5 Binäre Klassifikationsmodelle zur Vorhersage des Gewebeschicksals für eine personalisierte Therapieentscheidung 35

2.5.1 Validierung und Metriken 38

2.5.2 Algorithmen 42

2.5.3 Erklärbarkeit der Gewebeprognosen 49

2.5.4 Forschungsstand zur voxelbasierten Modellierung des Gewebeoutcomes 50

2.5.5 Forschungslücke 61

3 Vergleich zweier voxelbasierter Modellierungsansätze zur Integration räumlicher Informationen in die Vorhersage des Gewebeschicksals (lokal vs. global) 65

3.1 I-KNOW-Studie 65

3.1.1 Patienten 65

3.1.2 MRT-Protokoll 66

3.2 Berechnung der Bildfeatures für die Modellierung 67

3.3 MNI-Registrierung und Datenaufbereitung im MNI-Raum ... 73
3.4 Lokaler Ansatz ... 75
3.4.1 Anreicherung der lokalen Trainingsdaten ... 76
3.4.2 Erweiterung des lokalen Ansatzes ... 78
3.4.3 Globaler und hybrider Ansatz ... 79
3.5 Globaler Ansatz mit räumlichen Features ... 80
3.5.1 Räumliche Features ... 80
3.5.2 Modelltraining ... 81
3.6 Evaluierung ... 82
3.6.1 Validierung des lokalen Ansatzes ... 83
3.6.2 Validierung des globalen Ansatzes mit räumlichen Features ... 83
3.7 Feature Importance ... 83
3.7.1 Einfluss der Features bei der lokalen logistischen Regression ... 84
3.7.2 Einfluss der Features in XGBoost-Modellen ... 84
3.8 Statistische Analyse ... 87
3.8.1 Patientenstatistik ... 87
3.8.2 Modellstatistiken ... 87

4 Ergebnisse ... 89
4.1 Stichprobe ... 89
4.2 Lokaler Ansatz ... 90
4.2.1 Anreicherung der lokalen Trainingsdaten ... 90
4.2.2 Ergebnisse der Läsionsvorhersage ... 90
4.2.3 Einfluss der Features bei der lokalen logistischen Regression ... 94
4.3 Räumliche Features in globalen Modellen ... 94
4.3.1 Ergebnisse der Läsionsvorhersage ... 94
4.3.2 Einfluss der Features in XGBoost-Modellen ... 98
4.4 Vergleich zwischen lokalem Ansatz und globalem Ansatz mit räumlichen Features ... 104

5 Diskussion ... 107
5.1 Lokaler Ansatz ... 107
5.1.1 Vergleich der verschiedenen Modellvarianten ... 107
5.1.2 Erweiterung des lokalen Ansatzes ... 107
5.1.3 Anreicherung der Trainingsdaten ... 108
5.1.4 Skalierbarkeit ... 109
5.1.5 MNI-Registrierung ... 110
5.1.6 Implementierung weiterer Features ... 111
5.1.7 Zwischenfazit ... 112

5.2 Integration von räumlichen Features in den globalen Ansatz 114
5.2.1 Effekte auf die Läsionsvorhersagen 114
5.2.2 Räumliche Features in linearen Modellen 114
5.2.3 Räumliche Features in tree-basierten Modellen 115
5.2.4 Einfluss der Features in XGBoost-Modellen 117
5.2.5 Integration weiterer Features 118
5.2.6 Zwischenfazit 120
5.3 Vergleich beider Modellierungsansätze 122
5.3.1 Läsionsvorhersage 122
5.3.2 Erklärbarkeit 123
5.3.3 Limitationen 125
5.4 State of the Art 128

6 Ausblick 133

Anhang IX

Literaturverzeichnis XVII

Hinweise zur Veröffentlichung XXXI

Danksagung XXXIII

Curriculum Vitae XXXV

Abbildungsverzeichnis

Abbildung 1: Hauptursachen des Schlaganfalls 9
Abbildung 2: DWI-FLAIR-Mismatch 27
Abbildung 3: Time-of-Flight-Angiographie für einen proximalen rechtsseitigen Verschluss der mittleren Gehirnschlagader 30
Abbildung 4: Messprinzip der perfusionsgewichteten Bildgebung 32
Abbildung 5: PWI-DWI-Mismatch 34
Abbildung 6: Beispielhafte ROC-Kurven für drei fiktive Modelle 41
Abbildung 7: U-Net 48
Abbildung 8: Berechnung des ADC, Segmentierung des CSF und Unterteilung des Gehirns 68
Abbildung 9: PWI-Sequenz nach Korrektur von Patientenbewegungen für verschiedene Zeitpunkte 69
Abbildung 10: PWI-Sequenz nach Anwendung der Dekonvolution für verschiedene Zeitpunkte 70
Abbildung 11: Perfusionsparameter nach Dekonvolution und Normalisierung mittels kontralateraler Hemisphäre 71
Abbildung 12: Follow-up-FLAIR-Sequenz und Läsionsmaske 72
Abbildung 13: Beispiele für ausgeschlossene Patientendatensätze 73
Abbildung 14: Parameterkarten und Segmentierungen im MNI-Raum 74
Abbildung 15: Grundidee des lokalen Ansatzes 75
Abbildung 16: Räumliche Läsionsverteilung der I-KNOW-Patienten im MNI-Raum 77
Abbildung 17: Läsionsvorhersage beim lokalen Ansatz 93
Abbildung 18: Verteilung der Zielmetriken beim globalen Ansatz mit räumlichen Features 96
Abbildung 19: Vorhersagebeispiel beim globalen Ansatz mit räumlichen Features 97
Abbildung 20: Gain pro Feature für die besten XGB-Modelle 98
Abbildung 21: Durchschnittliche Shapley-Werte und Featureausprägungen pro Voxeloutcome 99
Abbildung 22: Wasserfallaufschlüsselung durchschnittlicher Shapley-Werte pro Voxeloutcome und Modell 100
Abbildung 23: Shapley-Wert-Zerlegung 101
Abbildung 24: Shapley-Wert-Zerlegungen 102
Abbildung 25: Permutation Importance pro Feature 103

Tabellenverzeichnis

Tabelle 1: Ausgewählte Veröffentlichungen zur Vorhersage des Gewebeoutcomes beim akuten ischämischen Schlaganfall 61

Tabelle 2: Eigenschaften der 99 Patienten aus der I-KNOW-Studie 89

Tabelle 3: Ergebnisse der Modelle beim lokalen und hybriden Ansatz 91

Tabelle 4: Ergebnisse der Modelle beim globalen Ansatz mit räumlichen Informationen 95

Tabellenverzeichnis (Anhang)

Tabelle A1: Hyperparametereinstellungen für XGBoost IX

Tabelle A2: Strukturierte Hypothesentests für die Modelle aus dem lokalen und hybriden Forschungsansatz X

Tabelle A3: Medianwerte der einzelnen Koeffizienten der lokalen logistischen Regression pro Gehirnregion XI

Tabelle A4: Durchschnittliche ROC AUCs und Dice-Koeffizienten für die globalen Modelle nach Integration räumlicher Features XII

Tabelle A5: Strukturierte Hypothesentests für die Modelle aus dem globalen Ansatz mit räumlichen Features XIII

Tabelle A6: Strukturierte Hypothesentests für den Vergleich der Modelle aus dem lokalen bzw. hybriden Ansatz und dem globalen Ansatz mit räumlichen Features XV

Abkürzungsverzeichnis

ADC	apparent diffusion coefficient [mm^2/s]
AHA	American Heart Association
AIF	arterial input function
ANOVA	analysis of variance
AnToNIa	Analysis Tool for Neuro Imaging Data
ANTs	Advanced Normalization Tools
ASA	American Stroke Association
ATP	Adenosintriphosphat
CBF	cerebral blood flow [ml/100 g/min]
CBV	cerebral blood volume [ml/100 g]
$CMRO_2$	cerebral metabolic rate of O_2 [ml/100 g/min]
CNN	convolutional neural network
CSF	cerebrospinal fluid
CT	Computertomographie
DALY	disability-adjusted life years
DEFUSE 2	Diffusion and Perfusion Imaging Evaluation for Understanding Stroke Evolution 2
DWI	diffusion-weighted imaging
ECASS	European Cooperative Acute Stroke Study
EPI	echoplanar imaging
ESCAPE	Endovascular treatment for Small Core and Anterior circulation Proximal occlusion with Emphasis on minimizing CT to recanalization times
EXTEND-IA	Extending the Time for Thrombolysis in Emergency Neurological Deficits – Intra-Arterial
FAST	face, arms, speech problems, time
FLAIR	Fluid-attenuated Inversion Recovery
FN	false-negative
FP	false-positive
FPR	False-Positive-Rate
FSL	FMRIB Software Library
FU	follow-up
GBD	Global Burden of Disease
HF	Hochfrequenz
I-KNOW	Integrating Information from Molecules to man: Knowledge Discovery Accelerates Drug Development and Personalized Treatment in Acute Stroke
iCAS	Imaging Collaterals in Acute Stroke
ISLES	Ischemic Stroke Lesion Segmentation
IV tPA	intravenous tissue plasminogen activator
KNN	künstliches neuronales Netz
LIME	local interpretable model-agnostic explanations
LOPO	Leave-One-Patient-Out
LP	lesion probability
LR	logistische Regression
MCA	middle cerebral artery
mmol	Millimol
MNI	Montreal Neurological Institute

MR-CLEAN	Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands
mRS	modified Rankin Scale
MRT	Magnetresonanztomographie
MTT	mean transit time [s]
NIHSS	National Institutes of Health Stroke Scale
PWI	perfusion-weighted imaging
RAPID	Rapid Processing of Perfusion and Diffusion
REVASCAT	Randomized Trial of Revascularization with Solitaire FR Device versus Best Medical Therapy in the Treatment of Acute Stroke Due to Anterior Circulation Large Vessel Occlusion Presenting within Eight Hours of Symptom Onset
RF	Random Forest
ROC AUC	area under the receiver operating characteristic curve
STAPLE	simultaneous truth and performance level estimation
SWIFT-PRIME	Solitaire™ With the Intention For Thrombectomy as PRIMary Endovascular Treatment
TE	echo time [s]
TIA	transitorisch ischämische Attacke
TICI	Thrombolysis in Cerebral Infarction
T_{max}	time to maximum of the residue function [s]
TN	true-negative
TP	true-positive
TPR	True-Positive-Rate
TR	repetition time [s]
TTP	time to peak [s]
UKE	Universitätsklinikum Hamburg-Eppendorf
WHO	World Health Organization
XGB	XGBoost
µm	Mikrometer

1 Einleitung

Der Schlaganfall ist weltweit die zweithäufigste Todesursache und eine der häufigsten Ursachen für körperliche und psychische Beeinträchtigungen [Katan und Luft, 2018]. Über 80 % der Schlaganfälle sind auf ein ganz oder teilweise verstopftes Blutgefäß zurückzuführen, das zu einer Ischämie, also Minderdurchblutung, im zu versorgenden Hirnareal führt [Schubert und Lalouschek, 2011]. Durch diese verminderte Durchblutung entsteht schnell ein Mangel an Sauerstoff, Glukose und anderen Nährstoffen, was den Stoffwechsel der Gehirnzellen beeinträchtigt. Je nach Grad und Dauer der Minderversorgung führt dies zu einem frühzeitigen Zelltod [Ermine et al., 2020]. Die ersten irreparabel geschädigten Zellen bilden den sogenannten Infarktkern, der sich bereits nach wenigen Minuten formt und sich dann auf den Rest des minderperfundierten Areals ausbreiten kann.

Das Ziel der akuten Schlaganfalltherapie besteht nun darin, dieses Gewebe, das auch als Penumbra (lat. Halbschatten) bezeichnet wird, durch eine rechtzeitige Rekanalisierung des verstopften Gefäßes zu retten [Catanese et al., 2017]. Da diese Minderversorgung in vielen Fällen durch kollaterale Umgehungskreisläufe zumindest partiell ausgeglichen wird, kann eine Penumbra noch viele Stunden nach Infarktbeginn vorhanden sein. Um irreparable großflächige Zellschäden abzuwenden, gilt es medizinisch schnell und angemessen zu reagieren [Campbell et al., 2013; Davis und Donnan, 2014].

Therapien zur Rekanalisierung wie die Thrombolyse und die Thrombektomie werden aufgrund ihrer Risiken nur in einem sehr engen Zeitfenster von 4,5 bzw. 6 Stunden nach ersten Symptomanzeichen allgemein empfohlen [Ringleb et al., 2021]. Nach Ablauf dieser Zeitspanne lässt sich zunächst eine Bewertung der Penumbra über spezifische medizinische Parameter aus einer zusätzlichen perfusions- und diffusionsgewichteten Bildgebung vornehmen [Cheng, 2021]. Zur Unterteilung des minderperfundierten Gewebes in Infarktkern und Penumbra wird in der Praxis routinemäßig das schwellwertbasierte Perfusions-Diffusions-Mismatch gebildet. Diese Art der Schwellwertbildung zur Vorhersage des Gewebeoutcomes vereinfacht die hohe Komplexität der zerebralen Perfusion und des Stoffwechsels sowie des dynamischen Läsionswachstums erheblich [Arakawa et al., 2006; Siemonsen et al., 2014]. Aus diesem Grund wurden in den letzten 20 Jahren vielfach binäre Klassifikationsmethoden zur Vorhersage der Penumbra und des Gewebeschicksals erprobt [Schlaug et al., 1999].

Obwohl diese Verfahren deutlich komplexere Zusammenhänge aus den Daten erlernen können, bleibt die Qualität der Prognosen bei alleiniger Nutzung von Perfusions- und Diffusionsparametern dennoch begrenzt, da viele weitere Faktoren wie u. a. die Qualität der Ground Truth, der Einbezug des Rekanalisationsstatus sowie physiologische Bedingungen eine Rolle spielen [Winzeck et al., 2018]. Dies gilt auch für künstliche neuronale Netze, für deren eigenständig aus den Bilddaten erlernten lokalen Muster üblicherweise die implizite Annahme getroffen wird, dass ein positionsunabhängiger Zusammenhang zwischen den regional erlernten Features und dem Gewebeschicksal besteht. Dieser Aspekt der Modellierung ist für den Schlaganfall allerdings inadäquat, da Läsionen durch arterielle Territorien eingeschränkt und nicht zufällig verteilt sind [Benzakoun et al., 2021]. Laut den beiden ISLES-Wettbewerben zur Vorhersage des Gewebeoutcomes sind deshalb klinische und a priori bekannte physiologische Informationen über Schlaganfall und Genesung zur Entwicklung von vielfältigeren Modellen einzubeziehen [Winzeck et al., 2018]. Bekannt ist zwar, dass Perfusionswerte innerhalb verschiedener Gewebearten und Gehirnregionen variieren [Payabvash et al., 2011] und sich auch die Qualität der kollateralen Durchblutung u. a. nach Lage des Gewebes unterscheidet [Vavilala et al., 2002; Sheth und Liebeskind, 2013; C. Chen et al., 2017], doch werden für die Vorhersage selten räumliche Faktoren berücksichtigt. Obwohl in früheren Studien bei Tieren [Shen und Duong, 2008] und Menschen [Kemmling et al., 2015] positive Effekte der Integration räumlicher Informationen nachgewiesen wurden, blieb eine detaillierte Verfolgung und Evaluierung dieses Ansatzes am Menschen bisher aus.

1.1 Zielsetzung

Die grundlegende Hypothese dieser Arbeit lautet, dass die Lage des Gewebes mit seinen anatomischen und physiologischen Informationen auf die Prädiktion des Gewebeschicksals einen deutlich positiven Effekt hat. Deshalb sollten neben den üblicherweise verwendeten Perfusions- und Diffusionsparametern auch räumliche Informationen in Modelle zur Prädiktion des Gewebeoutcomes integriert werden.

Die Hypothese soll anhand von zwei Forschungsansätzen mithilfe eines Datensatzes von 99 akuten Schlaganfallpatienten validiert werden:

1. Ansatz 1 basiert auf einer lokalen Modellierung und setzt sich aus je einem Modell pro Voxelposition zusammen. Bei diesem *lokalen Ansatz* wurde die Lage implizit verwendet. Zum Vergleich wurde ein Modell auf allen Voxelpositionen trai-

niert (*globaler Ansatz*). Dieses wurde ebenfalls verwendet, um den lokalen Ansatz an Positionen zu substituieren, an denen kein lokales Modell trainiert werden konnte. Zusätzlich wurde ein *hybrider Ansatz* als Mittelwert aus dem lokalen und globalen Ansatz gebildet, um ihre komplementären Vorhersagen zu vereinen. Koeffizienten der Modelle beim lokalen Ansatz wurden nach Möglichkeit auf Unterschiede bzgl. einzelner Gehirnregionen untersucht [Grosser et al., 2020b].

2. Für Ansatz 2 wurde ein globales Modell durch den expliziten Einbezug räumlicher Informationen in Form von Läsionswahrscheinlichkeitskarten und/oder Voxelpositionen erweitert. In einer vergleichenden Studie fand eine Analyse dieser Modelle hinsichtlich ihrer Berücksichtigung von räumlichen Informationen statt [Grosser et al., 2020a].

1.2 Aufbau der Arbeit

Nach der bisherigen inhaltlichen Hinführung in die Themenrelevanz werden im zweiten Kapitel die begrifflichen Grundlagen sowie der Forschungsstand zu Vorhersagen des Gewebeschicksals beim ischämischen Schlaganfall geklärt. Dazu wird in Abschnitt 2.1 zunächst das komplexe Krankheitsbild des akuten ischämischen Schlaganfalls mit seiner vielfältigen Symptomatik erläutert. Welche Instrumente der Diagnostik und Therapien zur Rekanalisierung eines verstopften Gefäßes derzeit bestehen, ist Inhalt der Abschnitte 2.2 und 2.3. Zur Diagnose des ischämischen Schlaganfalls kommen bildgebende Verfahren zum Einsatz, die in Abschnitt 2.4 dargelegt werden. Hier wird auf die Bestimmung der Penumbra durch die in der klinischen Praxis bislang routinemäßig eingesetzten Mismatch-Verfahren und deren Schwächen eingegangen. Daher sind in Abschnitt 2.5 binäre Klassifikationsmodelle und ihre Verwendung zur Vorhersage von infarziertem Gewebe beschrieben. In diesem Abschnitt wird u. a. Bezug auf gängige Konventionen bei der Modellierung, die wichtigsten Vorhersagealgorithmen sowie den aktuellen Stand der Forschung bzw. zu schließende Forschungslücken genommen.

Im dritten Kapitel werden zwei voxelbasierte Modellierungsansätze verglichen. In Abschnitt 3.1 werden zunächst die I-KNOW-Studie und die dort erhobenen Daten vorgestellt, die in dieser Arbeit zur Gegenüberstellung der beiden Forschungsansätze verwendet wurden. Im Anschluss daran sind in Abschnitt 3.2 die Bildverarbeitung und die Berechnung des scheinbaren Diffusionskoeffizienten und der Parameter aus der Perfusionsbildgebung (PWI, engl. perfusion-weighted imaging) beschrieben, die in allen Modellen dieser Arbeit vorkommen. Auf die weiteren Schritte zur Datenaufbereitung

für die Durchführung der Modellierungsansätze wird in Abschnitt 3.3 eingegangen. Eine konkrete Beschreibung der Ansätze erfolgt in den Abschnitten 3.4 (lokaler Ansatz) und 3.5 (globaler Ansatz mit räumlichen Features). In Abschnitt 3.6 sind die Evaluierungsschritte für die beiden Ansätze beschrieben. Zum besseren Verständnis der Auswirkungen von räumlichen Features sind die Auswertungen zur Feature Importance für die lokale logistische Regression und die globalen XGBoost-Modelle in Abschnitt 3.7 dargelegt. Hierzu werden insbesondere der sogenannte Gain, die Permutation Importance und Shapley-Werte eingeführt. Die statistische Analyse der in dieser Arbeit verwendeten Patientendaten findet sich in Abschnitt 3.8, in dem nicht nur die beiden Forschungsansätze, sondern auch deren Vergleich ausgeführt sind.

Im vierten Kapitel sind die Ergebnisse dieser Arbeit zusammengefasst: Zunächst werden in Abschnitt 4.1 die unterschiedlichen Resultate der statistischen Analyse der Patientenstichprobe im Hinblick auf die einzelnen I-KNOW-Studienstandorte angegeben, um danach die Ergebnisse des lokalen Ansatzes in Abschnitt 4.2 zu präsentieren. Dort wird als Erstes erklärt, inwiefern verschiedene Anreicherungsdesigns der lokalen Trainingsdaten die Anzahl der lokal trainierbaren Modelle beeinflussen. Die Prognosegüte des lokalen Ansatzes sowie Vergleiche zwischen dem hybriden und globalen Ansatz sind in Abschnitt 4.2.2 ausgewertet. Die Resultate zu den Koeffizienten der lokalen logistischen Regression pro Gehirnregion finden sich in Abschnitt 4.2.3. Anschließend werden die Ergebnisse der Läsionsvorhersage für die Modelle des globalen Ansatzes mit räumlichen Features in Abschnitt 4.3.1 dargestellt. Der Einfluss der einzelnen Features auf die besten XGBoost-Modelle (pro Featurekombination) wird in Abschnitt 4.3.2 ermittelt. Ein Vergleich zwischen den Modellen, die im Rahmen des ersten Forschungsansatzes trainiert wurden, und denen des globalen Ansatzes mit räumlichen Features findet sich in Abschnitt 4.4.

Im ersten Diskussionsteil sind zunächst die Vorhersageergebnisse des lokalen Ansatzes analysiert (Abschnitt 5.1.1), wobei der lokale Ansatz mit dem globalen (ohne räumliche Informationen) und dem hybriden Ansatz verglichen wird. Die Erweiterung des lokalen Ansatzes durch die Prädiktionen des globalen Modells stehen in Abschnitt 5.1.2 zur Diskussion. In Abschnitt 5.1.3 wird noch auf verschiedene Aspekte der Trainingsdaten der lokalen Modelle eingegangen. Die technischen Aspekte wie die Skalierbarkeit des lokalen Ansatzes und die für diesen Ansatz benötigte Registrierung der Bilddaten auf den Montreal Neurological Institute (MNI) Atlas sind in den Abschnitten 5.1.4 und 5.1.5 erörtert. Im nächsten Unterkapitel werden weiterführende Ideen zur

Implementierung zusätzlicher Features im Rahmen dieses Ansatzes präsentiert. Ein Zwischenfazit dazu, wie sich dieser Ansatz gewinnbringend in der Schlaganfalldiagnostik und Intervention verankern lässt, wird in Abschnitt 5.1.7 vorgenommen.

Im zweiten Teil der Diskussion steht die Analyse des globalen Ansatzes mit räumlichen Features an, sodass zunächst die Ergebnisse der Läsionsvorhersagen in Abschnitt 5.2.1 zusammengefasst sind. Die Nutzung der räumlichen Features durch die verschiedenen Algorithmen ist in den beiden Abschnitten 5.2.2–5.2.3 erläutert, wobei insbesondere der Einfluss der einzelnen Features auf die besten XGBoost-Modelle analysiert wird (vgl. Abschnitt 5.2.4). Neben zusätzlichen Ideen zur Integration weiterer Features wird resümiert, inwiefern sich räumliche Features sinnvoll in globale Modelle integrieren lassen (Abschnitte 5.2.5–5.2.6).

Nach der individuellen Betrachtung werden die beiden vorgestellten Forschungsansätze im dritten Diskussionsteil miteinander verglichen. Neben den unterschiedlichen Ergebnissen der Läsionsvorhersagen (Abschnitt 5.3.1) ist der Abschnitt 5.3.2 der Interpretierbarkeit von Vorhersagemodellen gewidmet, wobei auch die Limitationen dieser Arbeit genannt werden (Abschnitt 5.3.3). Die Befunde zur Läsionsvorhersage werden in Abschnitt 5.4 mit dem aktuellen empirischen Forschungsstand konfrontiert.

Im sechsten Kapitel schließt sich ein Ausblick auf die Entwicklungen von Vorhersagemodellen in Bezug auf die Ergebnisse dieser und anderer aktueller Arbeiten an. Nach einer kurzen Zusammenfassung der wichtigsten Erkenntnisse dieser Arbeit werden mittels einer Kontrastierung mit Resultaten aus anderen Studien Implikationen für die aktuelle Forschung herausgearbeitet. Zudem sind dort weitere Fragestellungen formuliert, die es in künftigen Studien zu klären gilt, um ein probates Tool zur Vorhersage von geschädigtem Gewebe beim ischämischen Schlaganfall zu entwickeln.

2 Diagnostische Verfahren zur Intervention beim akuten ischämischen Schlaganfall

2.1 Grundlagen

Nach der heute noch gültigen Definition der Weltgesundheitsorganisation (WHO, engl. World Health Organization) von 1970 bezeichnet der Schlaganfall „sich rasch entwickelnde Zeichen einer fokalen oder globalen Störung der zerebralen Funktion, woran sich Symptome anschließen, die 24 Stunden oder länger dauern oder gar zum Tode führen, ohne erkennbare Ursachen außer einer vaskulären“ [Aho et al., 1980, S. 114]. Diese Definition grenzt den Schlaganfall nicht nur von sich langsam entwickelnden Krankheitsbildern wie z. B. Hirntumoren ab, sondern vor allem von einer transitorisch ischämischen Attacke (TIA), deren Symptome sich nach WHO-Definition auf eine Dauer von 24 Stunden beschränken [Sacco et al., 2013].

Trotz dieser Einordnung und Spezifizierung des Schlaganfalls gegenüber anderen Krankheitsbildern lassen diese Definitionen außer Acht, dass auch bei stillen (nahezu symptomfreien) Infarkten und kurzer Symptomdauer bleibende Gehirnschäden entstehen können. Bei über einem Drittel der TIA-Patienten wurden akute Läsionen mittels diffusionsgewichteter Bildgebung (DWI, engl. diffusion-weighted imaging) nachgewiesen, woraus sich häufig irreversible Gehirnschäden entwickeln können [Easton et al., 2009]. Deshalb ist die eingangs genannte WHO-Begriffsbestimmung zum Schlaganfall in der Medizin nicht unumstritten, obwohl sie auch heute noch häufig Verwendung findet [Feigin et al., 2018].

Zu einer Änderung der TIA-Definition kam es 2009 durch die American Heart Association/American Stroke Association (AHA/ASA): Statt „ein plötzlicher, fokaler neurologischer Ausfall, der weniger als 24 Stunden andauert, vermutlich vaskulären Ursprungs ist und auf einen Bereich des Gehirns oder des Auges beschränkt ist, der von einer bestimmten Arterie durchblutet wird“ [Albers et al., 2002, S. 1715], wird nun eine TIA verstanden als „eine vorübergehende Episode einer neurologischen Funktionsstörung, verursacht durch eine fokale Ischämie des Gehirns, des Rückenmarks oder der Netzhaut, ohne akuten Infarkt“ [Easton et al., 2009, S. 2281].

Im Jahr 2013 erweiterte die AHA/ASA ihre Definitionen für den Schlaganfall noch hinsichtlich seiner Untertypen. Nun beinhaltet die Definition zusätzlich stille Infarkte und Gehirnblutungen, wobei die Zeitabhängigkeit durch einen radiologischen Nach-

weis ersetzt wurde. Sie unterscheidet sich demnach von derjenigen der WHO, weshalb die epidemiologischen Kennzahlen der AHA/ASA und der WHO zum Schlaganfall erheblich voneinander abweichen können [Feigin et al., 2018].

Unabhängig davon gilt bei einer akuten Symptomatik stets der Konsens, dass ein Schlaganfall als Ursache angenommen und schnellstmöglich mit geeigneten Diagnoseverfahren, wie sie u. a. in dieser Arbeit beschrieben werden, entschieden werden muss, welche Intervention angemessen ist. Zu den in der Praxis derzeit meist verwendeten Diagnoseinstrumenten gehören die bildgebenden Verfahren, die u. a. Aufschluss über den Gewebezustand des Gehirns geben, denn zu den Hauptursachen eines Schlaganfalls zählt die Minderdurchblutung des Gewebes (Ischämie; vgl. Abschnitt 2.1.1).

2.1.1 Ätiologie

Klassifiziert wird der akute Schlaganfall nach seinen Ursachen in den ischämischen und den hämorrhagischen Schlaganfall [Busch und Kuhnert, 2017]. Ca. 80 % der Schlaganfälle sind auf eine Minderdurchblutung des Gewebes (Ischämie) aufgrund eines akuten vollständigen oder partiellen Gefäßverschlusses zurückzuführen, der durch ein Blutgerinnsel entsteht [Schubert und Lalouschek, 2011; Busch und Kuhnert, 2017]. In ca. 15 % der Fälle ist die Ursache eine intrazerebrale Blutung (hämorrhagischer Schlaganfall), bei der nach einem Arterienriss Blut ins Gewebe strömt, wodurch die angrenzenden Hirnareale geschädigt werden und es zusätzlich zu einer Ischämie in nachgelagerten Arealen kommen kann [Schubert und Lalouschek, 2011]. In ca. 5 % der Fälle handelt es sich um eine Subarachnoidalblutung, die durch das Platzen einer Aussackung eines Gefäßes (Aneurysma) zwischen Gehirn und der weichen Hirnhaut (Subarachnoidalraum) hervorgerufen wird [Schubert und Lalouschek, 2011]. Die grundlegenden Ursachen eines Schlaganfalls sind in Abbildung 1 schematisch illustriert.

Die Ursache für einen ischämischen Infarkt liegt in 20–25 % der Fälle in einer Gefäßverengung der großen hirnversorgenden Gefäße wie beispielsweise der Halsschlagader (lat. Arteria carotis communis), die meist durch Gefäßverkalkung (Atherosklerose) ausgelöst wird. Durch Rupturen entzündlicher Stellen (atherosklerotische Plaques) an den Verengungen der inneren Gefäßwand entstehen Blutgerinnsel (Thromben), die das Gefäß noch weiter schmälern oder es sogar verschließen. Trotz Beeinträchtigung kann die Blutversorgung oftmals durch Umgehungskreisläufe ausrei-

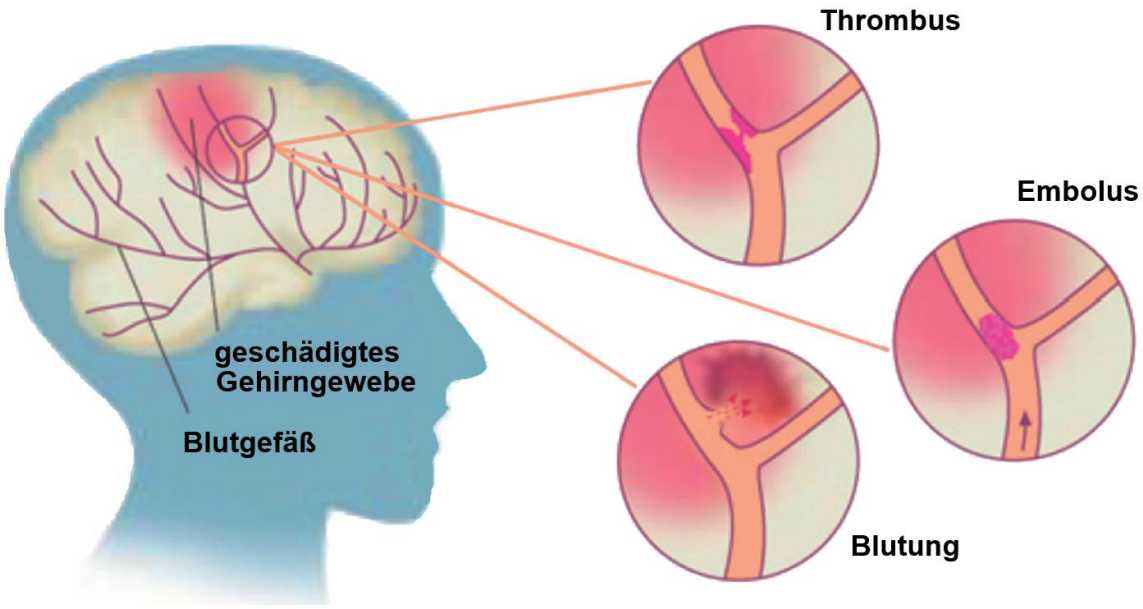

Abbildung 1: Hauptursachen des Schlaganfalls: Die Hauptursachen für einen Schlaganfall sind die Verstopfung eines Gefäßes durch einen Thrombus oder Embolus oder eine Gehirnblutung, bei welcher Blut frei ins Hirngewebe eindringt (Quelle: Bundesministerium für Bildung und Forschung, 2012).

chend aufrechterhalten werden. Wenn sich das Blutgerinnsel löst, wird es als sogenannter Embolus weitertransportiert. Ruft dieser andernorts eine Verstopfung hervor, spricht man von einer Embolie [Schubert und Lalouschek, 2011].

Ca. 20 % der ischämischen Schlaganfälle sind kleine (lakunäre) Infarkte, die beim Hirnstamm oder im Bereich der weißen Substanz[1] auftreten und durch den Verschluss eines kleinen hirnversorgenden Gefäßes entstehen [Schubert und Lalouschek, 2011]. In 20–25 % der ischämischen Schlaganfälle führen kleine Blutklumpen zu einem Verschluss in einem gehirnversorgenden Gefäß. Sie haben sich meist durch Vorhofflimmern oder Herzklappenerkrankungen in der linken Herzkammer gebildet und gelangen von dort in den Blutkreislauf (sogenannte kardiale Embolie) [Schubert und Lalouschek, 2011].

Zudem gibt es z. B. mit der Vaskulitis, bei der es aufgrund einer fehlerhaften Immunreaktion des Körpers zu einer Gefäßentzündung kommt, und der Thrombophilie, bei der das Blut eine erhöhte Gerinnungsaktivierung aufweist, einige seltenere Schlaganfallursachen, wobei sich in 20–40 % der Fälle auch keine bzw. keine eindeutige Ursa-

[1] Die weiße Substanz besteht im Wesentlichen aus Nervenzellfortsätzen (Axonen), welche für die Signalübertragung zuständig sind.

che feststellen lässt. Diese Schlaganfälle werden als kryptogene Schlaganfälle bezeichnet [Schubert und Lalouschek, 2011].

2.1.2 Epidemiologie

Um mehr über die Verbreitung von Schlaganfällen in verschiedenen Bevölkerungsgruppen zu erfahren, gibt es epidemiologische Studien, in denen verschiedene Kennzahlen des Krankheitsbilds ermittelt werden. So werden in der Global Burden of Disease-Studie (GBD) seit 1990 Ergebnisse zu mittlerweile über 300 Krankheiten und Unfällen, inkl. dem Schlaganfall, erfasst. In dieser weltweit umfassendsten epidemiologischen Beobachtungsstudie sind u. a. Alter und Geschlecht der betroffenen Personen sowie Jahr und Land der Vorfälle registriert. Bei Datenlücken werden auch transparente Schätzungen vorgenommen. Sofern in der vorliegenden Arbeit nichts anderes angegeben ist, beziehen sich die folgenden epidemiologischen Kennzahlen zum Schlaganfall auf die Kennzahlen der GBD-Studie für das Jahr 2019, die über das GBD-Online-Tool[2] abgerufen werden können.

Der Schlaganfall führt als weltweit zweithäufigste Ursache für Todesfälle und Behinderungen [Johnson et al., 2019; Krishnamurthi et al., 2020] dazu, dass sich die Gesamtzahl an vorzeitig verlorenen Lebensjahren durch Tod oder Behinderung (DALYs, engl. disability-adjusted life years) 2019 auf 143,2 Millionen Jahre beläuft.[3] 2019 sind weltweit 101,5 Millionen Menschen registriert, die schon einmal einen Schlaganfall hatten.[4] Davon erlitten 2019 etwa 12,2 Millionen Menschen einen ersten Infarkt.[5] Allein in Deutschland haben über 1,3 Millionen Menschen in ihrem Leben bereits einen Schlaganfall erlitten (Stand: 2019).[6] Dies entspricht bezogen auf die Gesamtbevölkerung Deutschlands einem Anteil von 1,76 % der Frauen und 1,34 % der Männer [Institute for Health Metrics and Evaluation (University of Washington), 2021]. Der Unterschied zwischen den Geschlechtern ist vor allem an der unterschiedlichen Lebenserwartung festzumachen. Das altersbereinigte Schlaganfallrisiko von Männern liegt in Deutschland etwa 1,2-mal so hoch wie das von Frauen [Appelros et al., 2009]. Insgesamt wur-

[2] http://ghdx.healthdata.org/gbd-results-tool
[3] Permalink: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/42115008a149f148cb278ac1d303ffc3
[4] Permalink: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/f80621f3c1103e19b091a6a744fb432b
[5] Permalink: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/806e3b2c0da4203950fb939a7022ce69
[6] Permalink: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/cbce8376bca5e374144a1af61ce20a43

den im Laufe des Jahres 2019 über 71.000 schlaganfallbedingte Todesfälle in Deutschland verzeichnet, was einer Sterblichkeitsrate von über 5 % entspricht.[7]

Für die Überlebenswahrscheinlichkeit und den Verlauf spielt die Schlaganfallursache eine wesentliche Rolle. Das akute Sterberisiko ist bei einer Gehirnblutung viermal so hoch wie bei einem ischämischen Infarkt. In der ersten Woche nach dem akuten Ereignis verringert sich das relative Sterberisiko und gleicht sich für beide Ursachen nach drei Monaten an. Adjustiert nach Alter, Geschlecht, Stärke des Infarkts und kardiovaskulären Risikofaktoren liegt das Risiko, bei einer Gehirnblutung zu sterben, 1,56-mal so hoch wie bei einem ischämischen Infarkt. Grundsätzlich machen Gehirnblutungen, gemessen an der skandinavischen Schlaganfallskala, nur 2 % der leichteren Infarkte, aber 30 % der schweren Infarkte aus [vgl. Andersen et al., 2009].

Für die Überlebenden bedeutet ein Schlaganfall häufig ein schweres gesundheitliches Schicksal: Sowohl die Art als auch das Ausmaß der Folgen hängen größtenteils davon ab, welche Gehirnregion wie stark geschädigt wurde. Etwa die Hälfte der Überlebenden wird nach einem Schlaganfall nicht wieder vollständig gesund. Ein Viertel benötigt Hilfe im täglichen Leben oder wird zum Teil pflegebedürftig bzw. schwerstbehindert [Truelsen et al., 2006; Kelly-Hayes, 2010]. Obwohl das individuelle Risiko, einen Schlaganfall zu erleiden, in den letzten Jahren zurückging, steigen die Gesamtzahlen aufgrund einer stetig wachsenden und immer älter werdenden Weltbevölkerung weiterhin an [Johnson et al., 2019].

Je nach Schwere der Erkrankung fallen Kosten für die Versorgung von Schlaganfallpatienten an, die sich in vier Kostenarten aufteilen lassen: Als erster Posten ist das Gesundheitswesen zu nennen (Primär-, Unfall- und Notversorgung sowie ambulante und stationäre Versorgung und Medikamente). Hinzu kommt zweitens die soziale Betreuung (Pflege und Heimunterbringung). Die beiden weiteren Posten bestehen aus Opportunitätskosten, die durch informelle Pflege über Angehörige entstehen, sowie Produktivitätskosten, die sich entweder aufgrund frühen Versterbens (Mortalität) oder Behinderung (Morbidität) ergeben. Nach Schätzungen sind durch die Versorgung von Schlaganfallpatienten im Jahr 2017 Kosten in Höhe von insgesamt 60 Milliarden Euro in Europa sowie in Island, Norwegen und der Schweiz, aber auch Israel angefallen.[8]

[7] Permalink: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/4fc9fd0ad11fcb374bdce6431ba66adc

[8] 8 % der weltweiten schlaganfallbedingten Todesfälle ereigneten sich 2017 in Europa, inkl. Island, Israel, Norwegen und der Schweiz.

Für Deutschland betrugen diese im selben Jahr 17,6 Milliarden Euro, was 0,54 % des Bruttoinlandproduktes entsprach [Luengo-Fernandez et al., 2019]. Aufgrund dieser hohen finanziellen Belastung des Gesundheitssystems, aber auch der einzelnen Länder, wird intensiv nach den Risikofaktoren geforscht, die einen Schlaganfall begünstigen.

2.1.3 Risikofaktoren

Bei den Risikofaktoren unterscheidet man zunächst, ob diese beeinflussbar sind oder nicht. Zu den nicht-beeinflussbaren Risikofaktoren zählen männliches Geschlecht, Ethnie sowie erhöhtes Alter, wie in Abschnitt 2.1.2 erwähnt [Kelly-Hayes, 2010; Boehme et al., 2017]. Das Risiko für einen Schlaganfall verdoppelt sich ab einem Alter von 45 Jahren etwa alle 10 Jahre [Kelly-Hayes, 2010].

Weitaus diffuser sind die beeinflussbaren Faktoren, denn dazu gehören u. a. Bluthochdruck, Diabetes mellitus, Rauchen, Vorhofflimmern, Stenosen der Halsschlagader, frühere Herzinfarkte und Schlaganfälle, übermäßiger Alkoholkonsum, erhöhte Cholesterinwerte[9], Übergewicht, Bewegungsmangel, wiederkehrende arterielle Claudicatio (umgangssprachlich auch als Schaufensterkrankheit bekannt), die Einnahme oraler Kontrazeptiva und anderer weiblicher Geschlechtshormone, chronische Infektionen und ein erhöhter Spiegel der Aminosäure Homozystein [Schubert und Lalouschek, 2011].

Zu der Frage, inwiefern die Risikofaktoren eher charakteristisch für ischämische Infarkte oder Gehirnblutungen sind und zu deren Differenzierbarkeit beitragen, liegen noch keine konsistenten Ergebnisse aus der medizinischen Forschung vor. Bislang geht man bei ischämischen Anfällen davon aus, dass diese durch Risikofaktoren für atherosklerotische Herz-Kreislauf-Erkrankungen wie Diabetes, Vorhofflimmern, bisherige Herzinfarkte und Schlaganfälle insbesondere begünstigt werden [Andersen et al., 2009]. Zudem nimmt lange bestehender Bluthochdruck nach bisherigen Erkenntnissen eine besondere Rolle bei Gehirnblutungen ein, obwohl dieser ursächlich für beide Infarkte genannt wird [Schubert und Lalouschek, 2011]. Während leichter Alkoholkonsum einen präventiven Einfluss auf ischämische Infarkte zu haben scheint, ist die Studienlage zum Einfluss von Rauchen und übermäßigem Alkoholkonsum auf die unterschiedlichen Schlaganfalltypen weniger eindeutig [Andersen et al., 2009].

[9] Diese sind beim Schlaganfall allerdings weniger relevant als beim Herzinfarkt.

Obwohl die Ursachen für einen Schlaganfall ungeklärt und die Risikofaktoren vielfältig und multikausal sein können, lässt sich ein ischämischer Infarkt eindeutiger bestimmen: Als primärer Indikator für die möglichen Folgen lässt sich die Beeinträchtigung der Blutversorgung im Gehirn messen.

2.1.4 Die arterielle Blutversorgung des Gehirns

Das Gehirn benötigt für seinen Stoffwechsel etwa 20 % der Energie des Herzzeitvolumens, obwohl das Organ nur etwa 2 % der Körpermasse ausmacht [Reich und Nikoubashman, 2016]. Die sogenannte graue Substanz (hauptsächlich Zellkörper) verbraucht etwa dreimal so viel Sauerstoff wie die weiße Substanz. Neben Sauerstoff benötigen die Zellen vor allem Glukose, die als hauptsächliches Substrat zur Erzeugung von Adenosintriphosphat (ATP) dient: der Hauptenergiequelle der Zellen.[10] Da das Gehirn kaum über Kapazität zur Speicherung von Energiesubstraten verfügt, ist es auf eine konstante Durchblutung zur Versorgung mit Sauerstoff und Glukose angewiesen [Vavilala et al., 2002].[11] Bei einer Unterversorgung durch einen gestörten Blutfluss fehlt es an Sauerstoff zur ATP-Produktion und damit an Energie für den Stoffwechselprozess der Zellen. Normalerweise wird durch die Natrium-Kalium-ATP-Pumpe ein kontinuierlicher Transport von Natrium-Ionen aus der Zelle und Kalium-Ionen in die Zelle aufrechterhalten. Es wird vermutet, dass der ATP-Mangel zu einer veränderten Membrandurchlässigkeit führt und extrazelluläres Wasser in die Zellen gelangt [Moseley et al., 1990]. Sofern der Blutfluss vor dem Zusammenbruch der Zellmembran wiederhergestellt werden kann, sind die Schäden jedoch reversibel [Brown und Semelka, 2003].

Dieser hohe Bedarf hat zur Entwicklung spezifischer zerebraler Blutgefäßnetze mit arteriovenöser Hierarchie geführt [Andjelkovic et al., 2020]. Die Blutversorgung des Gehirns erfolgt vorne über die rechte und linke innere Halsschlagader (lat. Arteriae carotides internae) und hinten über die Wirbelarterien (lat. Arteriae vertebrales). Beide Arterienpaare sind jeweils über Anastomosen untereinander vernetzt, sodass Umgehungskreisläufe existieren, die den Ausfall eines Gefäßes temporär ausgleichen können. Die inneren Halsschlagadern münden in den vorderen Teil des Circulus Willisi.

[10] Im Ruhezustand resultiert bis zu 92 % des ATPs aus dem oxidativen Metabolismus von Glukose. Ein kleiner Teil wird über Laktat gedeckt.
[11] Die ATP-Reserven des Gehirns sind ohne Sauerstoffzufuhr nach etwa 7 Minuten aufgebraucht.

Die Wirbelarterien gehen in die Arteria basilaris über, welche den hinteren Teil des Großhirns, einen Teil des Kleinhirns und den Hirnstamm versorgt, und zusammen mit den inneren Halsschlagadern den Circulus Willisi formt [Andjelkovic et al., 2020]. Allerdings muss berücksichtigt werden, dass die genaue Ausprägung des Circulus Willisi häufig individuell variiert. Bei etwa 70 % der Bevölkerung sind Teile dieses Arteriensytems nicht vollständig ausgebildet oder gar nicht vorhanden [Van Kammen et al., 2018]. Dieses ringförmige als Anastomose fungierende Arteriensystem am Mittelhirn teilt das vordere Stromgebiet in die drei Hauptgefäße zur Versorgung des Groß- und des Zwischenhirns auf: die vordere (lat. Arteria cerebri anterior), die mittlere (MCA, engl. middle cerebral artery, lat. Arteria cerebri media) und die hintere (lat. Arteria cerebri posterior) Gehirnschlagader [Vavilala et al., 2002]. Jedes der drei Hauptgefäße zerteilt sich weiter in kleinere Gefäße und Arteriolen, deren Äste teilweise über die leptomeningealen Anastomosen verbunden sind und an der Hirnoberfläche verlaufen bevor sie ins Hirnparenchym eindringen [Vavilala et al., 2002; Andjelkovic et al., 2020].

Die von den Hauptästen der inneren Halsschlagader versorgten Areale haben mit Ausnahme des Stromgebiets der mittleren Gehirnschlagader eine gute kollaterale Durchblutung. Deshalb ist ihr Stromgebiet besonders anfällig für Ischämien [Vavilala et al., 2002]. Kollateralen unterscheiden sich von Anastomosen darin, dass sie keine verschiedenen Gefäßgebiete, sondern Seitenäste einer Arterie miteinander verbinden. Damit können potenzielle Gefäßverschlüsse überbrückt werden [Sesto, 2013]. Die Qualität der kollateralen Durchblutung ist neben der Lage eines Verschlusses ein aussagekräftiger Prädiktor für das klinische Outcome der Schlaganfallpatienten, da bei einem Verschluss das gesamte von der Arterie zu versorgende Territorium bedroht ist [Saarinen et al., 2014].

Der Gefäßbaum verzweigt sich nach den Hauptästen weiter in Arteriolen und winzige Kapillaren, welche eine netzartige Struktur bilden. Die Fläche des Gefäßbettes vergrößert sich dramatisch, wodurch das Blut langsamer wird und die Diffusion von Sauerstoff und andere kapillare Austauschvorgänge ermöglicht werden. Die netzartige Struktur minimiert außerdem die Effekte von Verschlüssen innerhalb der Kapillaren. Die Dichte der Kapillaren ist dabei so dünn, dass der Großteil der Gehirnzellen weniger als 25 µm von einem Gefäß entfernt ist, wobei sich die Anzahl der Gefäße nach Gehirnregion und Gewebetyp unterscheidet. Der Stoffaustausch mit dem Gewebe durch die als Blut-Hirn-Schranke bezeichnete Gefäßummantelung erfolgt hochselektiv. Der

Abtransport der von den Kapillaren aufgenommenen Stoffe erfolgt über die Venolen, die venösen Entsprechungen der Arteriolen [Andjelkovic et al., 2020].

Außer dem geschädigten Nervengewebe im Gehirn existieren noch weitere, teilweise sichtbare Symptome beim akuten ischämischen Schlaganfall. Diese werden im nächsten Abschnitt aufgeführt.

2.2 Akute Symptomatik und Gewebeschäden

Zu den häufigsten Schlaganfallsymptomen gehören – je nach geschädigtem Areal – halbseitige Lähmungserscheinungen oder Gefühlsstörungen, Sprachstörungen oder bestimmte Formen von Sehstörungen [Schubert und Lalouschek, 2011]. Hierzu zählen u. a. eine plötzlich auftretende einseitige Schwäche, Taubheit oder Sehverlust, doppeltes Sehen (Diplopie), veränderte Sprache, Bewegungseinschränkungen (Ataxie) und nicht-orthostatischer Schwindel (ohne Stehbelastung). Als Begleiterscheinungen sind starke Kopfschmerzen zu nennen, die aus den Infarktursachen folgen. Untypische Symptome sind u. a. isolierter Schwindel, beidseitige Blindheit, Amnesie, Probleme beim Steuern und Ausführen von Sprachbewegungen (Dysarthrie), Schluckstörungen (Dysphagie), Atemgeräusche aufgrund von Verengungen der Atemwege (Stridor), Fremdakzent oder Kopfschmerzen, unkontrollierte Bewegungen eines Arms oder Beins (Hemiballismus), Verlust der Kontrolle über eine der Hände (Alien-Hand-Syndrom), Verwirrtheit und verändertes Bewusstsein sowie eine fehlende Wahrnehmung zur Erkennung der Symptome (Anosognosie) [Hankey, 2017].

Als einfaches und sensitives Diagnosehilfsmittel zur frühzeitigen Erkennung für Ersthelfer gilt u. a. der FAST-Test [Kleindorfer et al., 2007; Hankey, 2017]. Die ersten drei Buchstaben dieses Akronyms stehen für Fragen an den Betroffenen zu den Bereichen Face (*Kann die Person lächeln? Ist dabei eine Seite gelähmt?*), Arms (*Kann die Person beide Arme hochheben und oben halten?*) und Speech Problems (*Kann die Person klar und deutlich sprechen?*). Das T steht für Time, da es sich bei der Bejahung einer der vorherigen Fragen mit hoher Wahrscheinlichkeit um einen Schlaganfall handelt und so schnell wie möglich der Rettungsdienst gerufen werden muss [Kleindorfer et al., 2007; Dombrowski et al., 2014].

Alle akuten Schlaganfallpatienten sind auf einer Stroke Unit zu behandeln, auf der ein engmaschiges apparatives Monitoring mit regelmäßiger Kontrolle der Vitalparameter und eine häufige klinische Untersuchung erfolgt [Ringleb et al., 2021]. Diese Sta-

tionen sind eigens auf die Therapie von akuten Schlaganfallpatienten spezialisiert und bieten sowohl Diagnose- als auch Therapiemöglichkeiten. Zumeist sind sie einer neurologischen Abteilung angegliedert [Faiss et al., 2008]. Dort erfolgt bei allen Patienten mit Verdacht auf Schlaganfall eine Untersuchung mittels bildgebenden Verfahren, wenn für diese eine Reperfusionstherapie infrage kommt, um den Gewebeschaden aufzuhalten [Ringleb et al., 2021]. Hierzu wird die Computertomographie (CT) und/oder die Magnetresonanztomographie (MRT) eingesetzt, die in Abschnitt 2.4 genauer vorgestellt wird. Für die weitere Behandlung braucht es Erkenntnisse zur Schlaganfallursache und die Gewissheit, dass eine Gehirnblutung ausgeschlossen werden kann [Berkefeld und Neumann-Haefelin, 2009].

Ein Indikator für eine Schlaganfalldiagnose kann auch eine Schlaganfallmimik sein, die immerhin etwa 20–25 % der Schlaganfallpatienten aufweisen. Besonders schwierig ist die Diagnose außerdem in Fällen von atypischen oder wechselnden Symptomen, Unwohlsein oder Verwirrtheit der Patienten. Darüber hinaus erschwert ein verspäteter Zugang zur Bildgebung eine optimale Behandlung [Hankey, 2017].

Zur Messung des Schweregrades der Symptome bietet sich die *NIHSS* (National Institutes of Health Stroke Scale) an. Diese einheitliche Methodik wird auch in Studien verwendet, um eine Vergleichsbasis zwischen den Infarktauswirkungen nahe am Ereignis und denjenigen an Folgezeitpunkten zu haben. Dies dient als Bewertungsgrundlage, welche Maßnahmen zur Verbesserung des Outcomes einzuleiten sind, oder wie sich das finale Outcome mit einer Baseline adjustieren lässt [Lyden, 2017].

Die NIHSS wurde so konzipiert, dass geschultes Personal alle Tests mit dem Patienten direkt am Krankenbett durchführen kann [Ortiz und Sacco, 2007]. Je nach Variante[12] umfasst die Skala elf Items mit Werten von null (keine Beeinträchtigung) bis maximal vier (schwerste Beeinträchtigung). Sie gewichtet linksseitige Infarkte um etwa vier Skalenpunkte stärker als rechtsseitige, da sie u. a. mehrere Items zur Sprache und Koordination beinhaltet [Woo et al., 1999; Lyden, 2017]. Dafür weist sie eine annehmbare interne Konsistenz (Cronbachs Alpha) sowie eine hohe Reproduzierbarkeit auf. Ihre Gesamtsumme korreliert mit dem Infarktvolumen und ist damit ein guter Prädiktor für das Infarktoutcome und die Präsenz großer Arterienverschlüsse [Lyden, 2017].

[12] Die originale Cincinnati/Naloxone NIHSS Variante umfasste 15 Items, die einen maximalen Score von 42 ergaben.

Obwohl das Blutungsrisiko bei einer Thrombolyse mit der Größe des Schlaganfalls zunimmt, ist der NIHSS nicht gleichzusetzen mit einem Behinderungsgrad. Für Patienten mit Behinderungen sollte sowohl bei leichten (NIHSS ≤ 5) als auch schweren klinischen Symptomen (NIHSS ≥ 25) nach Möglichkeit eine Thrombolyse durchgeführt werden (vgl. Abschnitt 2.3) [Ringleb et al., 2021].

Zur Quantifizierung des funktionellen Outcomes wird – vor allem in klinischen Studien – die sogenannte *modifizierte Rankin-Skala* (mRS) verwendet [Quinn et al., 2009]. Dieses Scoring-System besteht aus den folgenden Abstufungen, um die Auswirkungen des Schlaganfalls zu klassifizieren [Van Swieten et al., 1988]:

0 – keine Symptome
1 – keine wesentliche Beeinträchtigung: kann trotz gewisser Symptome Alltagsaktivitäten ausüben.
2 – leichte Beeinträchtigung: ist im Alltag eingeschränkt, kann sich aber ohne Hilfe versorgen.
3 – mittelschwere Beeinträchtigung: ist im Alltag eingeschränkt, kann aber ohne Hilfe gehen.
4 – höhergradige Beeinträchtigung: kann nicht ohne Hilfe gehen, benötigt Hilfe bei der Körperpflege.
5 – schwere Behinderung: bettlägerig, inkontinent und benötigt ständige pflegerische Unterstützung.

Für klinische Studien wird die Skala für verstorbene Patienten häufig um den Wert 6 ergänzt [Quinn et al., 2009].

Trotz der bildgebenden Verfahren und Skalen zur Diagnose eines Schlaganfalls hinsichtlich des Schweregrads mangelt es an Instrumenten, die den Gewebeschaden im Gehirn präzise prognostizieren können. Und das, obwohl dieser die verlässlichste Symptomatik eines akuten ischämischen Schlaganfalls darstellt und damit die sicherste Entscheidungsgrundlage für die Auswahl an akuten Therapien zur Rettung des Gewebes gewährleistet. Gelingt die Einschätzung des Gewebeschadens, dann lässt sich die Therapie beim akuten ischämischen Schlaganfall besser abstimmen. Welche Therapiemöglichkeiten derzeit beim akuten ischämischen Schlaganfall bestehen, ist im nächsten Kapitel erläutert.

2.3 Rekanalisationstherapien zur Rettung der Penumbra

Die primäre Zielsetzung der akuten Schlaganfallbehandlung besteht darin, das minderperfundierte, jedoch noch nicht irreparabel geschädigte Gebiet im Gehirn, die Penumbra wiederherzustellen [Catanese et al., 2017]. Zwei bewährte Therapieoptionen, die beim akuten ischämischen Infarkt eingesetzt werden, um die ursächliche Verstopfung des Gefäßes aufzulösen und den Blutfluss in den betroffenen Gehirnregionen wiederherzustellen, sind die *Thrombolyse* und die *Thrombektomie* [El Tawil und Muir, 2017].

Bei der Thrombolyse wird ein Gerinnsel auflösendes Medikament injiziert, wodurch das verstopfte Gefäß mit einer 30-prozentigen Wahrscheinlichkeit rekanalisiert wird [Baron, 2018]. Die Chance der Rekanalisierung hängt dabei von der Länge [Riedel et al., 2011] und der Lage der Verstopfung ab [Saarinen et al., 2012]. Zuvor muss jedoch eine Gehirnblutung aufgrund der Kontraindikation mittels MRT oder CT ausgeschlossen werden [Berkefeld und Neumann-Haefelin, 2009].

Das Zeitfenster, in dem eine Thrombolyse üblicherweise durchgeführt wird, wurde zwar von der AHA/ASA 2009 nach den Ergebnissen aus ECASS (European Cooperative Acute Stroke Study) III aus dem Jahr 2008 von 3 Stunden auf bis zu 4,5 Stunden nach Beginn des Infarkts ausgeweitet [Del Zoppo et al., 2009]. Dennoch muss einer Thrombolyse nach Ablauf dieses Zeitfensters eine sorgfältige Risiko-Nutzen-Abwägung vorausgehen, weil sich das Risiko einer Gehirnblutung und ihrer u. a. lebensbedrohlichen Konsequenzen erhöht [Cossey und Gonzales, 2015]. Deshalb wird zumeist nur eine Thrombolyse für Patienten im unklaren Zeitfenster und im 4,5- bis 9-Stunden-Zeitfenster empfohlen, sofern die Patienten ein Mismatch (MRT: DWI-FLAIR oder PWI-DWI, CT: PWI) aufweisen und sonstige Einschlusskriterien erfüllen [Ringleb et al., 2021].

Eine weitere Behandlungsmethode zur Rekanalisierung etablierte sich vor wenigen Jahren mit der Thrombektomie. Bei dieser Methode wird das Blutgerinnsel mit einem mechanischen Eingriff aus dem Gefäß entfernt. Nach zunächst widersprüchlicher Studienlage konnten mehrere Studien (MR-CLEAN, ESCAPE, EXTEND-IA, SWIFT-PRIME, REVASCAT) die bis 2015 bestehenden Zweifel an dieser Methode ausräumen. Infolgedessen wurde die Thrombektomie weltweit in die Leitlinien für die akute Behandlung des ischämischen Schlaganfalls aufgenommen [Friedrich et al., 2020], da

auch die Rekanalisierungsrate bei dieser Behandlungsmethode mit über 80 % weit über derjenigen der Thrombolyse liegt [Baron, 2018].

Das hierzu benötigte Wissen über die Lage des Verschlusses wird häufig mithilfe kranialer Angiographie abgeklärt [J.-T. Kim et al., 2019]. Bei einer Thrombektomie wird der Patient oft nur lokal betäubt und bleibt ansprechbar. Das genaue Prozedere variiert je nach Art, Ursprung und Lage des Verschlusses. Ein Mikrokatheter wird entlang eines Führungskatheters über die Leistenarterie eingeführt und bis ins verstopfte Gefäß geleitet, wo dieser durch das Gerinnsel geschoben wird. Nun wird ein Stent-Retriever durch den Mikrokatheter gelegt. Nach leichtem Zurückziehen des Katheters liegt das Gitternetz des Retrievers (mit einem Überschuss auf beiden Seiten) inmitten des Blutgerinnsels, in dem es sich optimalerweise verfängt. Beim Zurückziehen des Stents mithilfe von Unterdruck wird das Blutgerinnsel freigesetzt und letztendlich entfernt [Evans et al., 2017]. Teilweise sind mehrere Versuche für eine erfolgreiche Rekanalisation nötig [Flottmann et al., 2018].[13]

Der Rekanalisationsstatus lässt sich anhand des *TICI-Scores* (Thrombolysis in Cerebral Infarction) klassifizieren [Higashida und Furlan, 2003]:

0 – keine nach-vorne-gerichtete Perfusion jenseits des Verschlusses.
1 – Penetration mit minimaler Perfusion: Das Kontrastmittel passiert den Verschluss füllt jedoch während der angiographischen Untersuchung nur einen kleinen Teil des Gefäßbetts.
2 – Partielle Perfusion:
    2a – weniger als 2/3 des betroffenen Gebiets sind wiederdurchblutet.
    2b – vollständige, aber verlangsamte Perfusion.
3 – vollständige Perfusion.

Als verhältnismäßig sicher gilt die Thrombektomie, wenn diese von geübtem Fachpersonal durchgeführt wird – trotz möglicher Komplikationen wie vom Device verursachten Gefäßperforationen, Gefäßdissektionen und Subarachnoidalblutungen, Schäden beim Legen des Gefäßzugangs sowie Problemen mit dem Kontrastmittel [Evans et al., 2017]. Auch die mechanische Thrombektomie hat bei Verschlüssen der großen Arterien im vorderen Stromgebiet so schnell wie möglich und innerhalb der ersten 6 Stunden nach Infarktbeginn zu erfolgen. Nach diesem Zeitfenster wird nur noch dann

[13] Dieser Schritt ist kritisch, da Verletzungen an den Gefäßen beim Entfernen des Thrombus entstehen können [Evans et al., 2017].

zur Thrombektomie geraten, wenn sich die Existenz einer Penumbra auf Basis der Bildgebung vermuten lässt [Ringleb et al., 2021].

Während sich die Thrombolyse vor allem für die Behandlung von Verschlüssen kleinerer Gefäße eignet, wird die Thrombektomie vor allem bei Verschlüssen der größeren Gefäße im vorderen Stromgebiet angewandt [Jayaraman et al., 2017]. Für welche der beiden Therapieverfahren sich letztendlich medizinisch entschieden wird, hängt von den betroffenen Gefäßen und der Entwicklung des Gewebeschadens ab. Zur Identifikation braucht es Verfahren der medizinischen Bildgebung, die im nächsten Kapitel erläutert werden.

2.4 Bildgebung mittels Magnetresonanztomographie

Bei Verdacht auf einen Schlaganfall muss so schnell wie möglich eine kraniale medizinische Bildgebung erfolgen, um vor dem Einsatz einer möglichen Reperfusionstherapie eine Gehirnblutung auszuschließen [Schubert und Lalouschek, 2011; Ringleb et al., 2021]. Einerseits dienen die bildgebenden Verfahren dazu, die Diagnose und den Ausschluss einer Blutung zu bestätigen. Andererseits werden die weiteren Interventionsmaßnahmen durch das Feststellen eines akuten ischämischen Schlaganfalls bestimmt, um Gewebeschäden zu verhindern. In der Praxis dienen hierzu die CT und die MRT. Als erste Instanz in der akuten Schlaganfall-Diagnostik gilt weltweit die CT, aufgrund ihrer hohen Verfügbarkeit und geringen Untersuchungsdauer. Die CT wird wesentlich häufiger eingesetzt als die MRT [Vilela und Rowley, 2017]. Beide bieten jedoch die Möglichkeit, ein vollumfängliches Schlaganfallprotokoll durchzuführen, wobei dann die zeitlichen Unterschiede für diese beiden Verfahren nicht mehr so stark ausfallen [Thomalla et al., 2009].

Während ein Blutungsausschluss bei der CT mithilfe nativer (kontrastmittelfreier) CT vorgenommen wird, geschieht dies bei der MRT z. B. mit der T2*-gewichteten Sequenz [Hankey, 2017; Vert et al., 2017]. Für Patienten, die nach einem Blutungsausschluss für eine Thrombektomie infrage kommen, wird im Rahmen einer Angiographie zunächst der Zustand der Gefäße untersucht und der Ort des Verschlusses detektiert, um den Verschluss zielgerichtet zu entfernen [Ringleb et al., 2021].

Ist der Zeitpunkt der ersten Symptomanzeichen unbekannt oder liegen diese bereits länger als 4,5 (Thrombolyse) bzw. 6 Stunden (Thrombektomie) zurück, muss zunächst die Existenz und der Umfang von potenziell rettbarem Gewebe festgestellt werden,

welches den Einsatz der Therapien in Anbetracht eines erhöhten Blutungsrisikos rechtfertigt [Ringleb et al., 2021]. Dies betrifft immerhin knapp 75 % der Schlaganfallpatienten, für die eine individuelle Risikobewertung vorgenommen werden muss, da sie verspätet in der Klinik eintreffen [Tong et al., 2012].

Die Quantifizierung einer Penumbra aus der Differenz der Infarktprognose und dem aktuellen Gewebeschaden (Infarktkern) basiert in der klinischen Praxis routinemäßig auf verschiedenen Mismatch-Verfahren. Während bei der CT die Einschätzung der zukünftigen und der aktuellen Infarktregion jeweils anhand der PWI-Bildgebung vorgenommen wird, steht bei der MRT zusätzlich die DWI-Bildgebung zur Verfügung, welche einen Infarkt frühzeitig anzeigt und dort zur Identifikation des Infarktkerns dient [Schellinger et al., 2010].

Da die MRT mit ihrem DWI-FLAIR-Mismatch u. a. eine Beurteilung des Infarktalters bei Patienten mit unklarem Infarktzeitpunkt erlaubt, ist sie besonders im späten Zeitfenster relevant [Ringleb et al., 2021].

Weil die Daten aus der I-KNOW-Studie mithilfe der MRT erhoben wurden, werden nach der folgenden Einführung in die MRT die Sequenzen aus dem Schlaganfallprotokoll der I-KNOW-Studie erläutert. Diese Daten fließen im weiteren Verlauf dieser Arbeit in die Modellierung des Gewebeoutcomes mithilfe von binären Klassifikationsmodellen ein.

2.4.1 Mechanismen der Magnetresonanztomographie

Die Magnetresonanztomographie (MRT) bietet ein hochauflösendes Verfahren zur nichtinvasiven schichtweisen Darstellung von Körpergewebe. Im Gegensatz zur CT kommt die MRT ohne ionisierende Röntgenstrahlung aus, die mit dem Risiko einhergeht, kumuliert in höheren Dosen Krebs auslösen zu können. Zudem ist besonders ihr guter Weichteilkontrast positiv in Bezug auf den Schlaganfall gegenüber der CT hervorzuheben, was sie u. a. sensitiver für kleinere Infarkte macht. Darüber hinaus bietet sie mit der diffusionsgewichteten Bildgebung die Möglichkeit, Infarkte besser in einem frühen Stadium zu erkennen. Jedoch eignet sich eine MRT-Untersuchung, aufgrund des starken Magnetfeldes, nicht für Patienten mit beispielsweise Herzschrittmachern, obwohl Metallimplantate bei beiden Verfahren grundsätzlich zu Artefakten führen können [Berkefeld und Neumann-Haefelin, 2009; Thomalla et al., 2009].

Die Mechanismen der MRT werden in diesem Abschnitt auf Basis der Ausführungen von [Brown und Semelka, 2003] sowie in [Siemens Medical, 2003] beschrieben. Anschließend werden die wichtigsten Sequenzen und Techniken eines multimodalen MRT-Protokolls beim Schlaganfall erläutert, die auch bei der Untersuchung der Patienten aus der I-KNOW-Studie angewendet wurden: die Fluid-attenuated Inversion Recovery (FLAIR), die Time-of-Flight-MR-Angiographie, die DWI- und die PWI-MRT-Bildgebung [Vert et al., 2017].

Ausgangspunkt für jede MRT-Messung ist stets ein äußeres Magnetfeld, unter dessen Einfluss die Kernspins der einzelnen Atome des Messkörpers eine Gleichgewichtsmagnetisierung erzeugen.

Nettomagnetisierung

Atomkerne bestehen aus Protonen und Neutronen, die jeweils einen als Spin bezeichneten Eigendrehimpuls besitzen. Da sich die Spins identischer Kernteilchen gegenseitig aufheben, besitzen nur Atomkerne mit einer ungeraden Anzahl an Protonen und/oder Neutronen einen sogenannten Kernspin. Atome mit einem Kernspin produzieren ein als magnetisches Dipolmoment bezeichnetes lokales Magnetfeld, indem sich die Spins mit konstanter Frequenz um ihre Drehachse bewegen (präzessieren). Die MRT basiert auf der Anwendung eines statischen äußeren Magnetfeldes, entlang dessen sich die einzelnen Spinmagnete ausrichten.

Da für nahezu jedes Element aus dem Periodensystem Isotope[14] mit einem Kernspin existieren, lässt sich prinzipiell fast jeder Stoff mithilfe der MRT untersuchen. Ein einfaches Beispiel für einen Atomkern mit einem Spin ist Protium (^{1}H), das in diesem Kontext von besonderer Bedeutung ist, da Wasserstoff einen Großteil des menschlichen (Gehirn-)Gewebes ausmacht. Es ist zugleich das am häufigsten vorkommende Wasserstoffisotop. Viele MR-Verfahren zielen darauf ab, die Anzahl von ^{1}H-Atomen innerhalb eines Volumenelements (auch Voxel genannt) zu messen, da pathologisches Gewebe eine andere Wasserkonzentration als das umgebende Gewebe aufweist.

Ohne äußeres Magnetfeld sind die lokalen Magnetfelder der ^{1}H-Atomkerne rein zufällig orientiert und kompensieren sich gegenseitig. Unter Anwendung eines äußeren

[14] Isotope bezeichnen eine Gruppe von Atomen mit gleich vielen Protonen, jedoch unterschiedlich vielen Neutronen.

magnetischen Feldes B_0 richten sich die Spinmagnete parallel oder entgegengesetzt zu diesem aus. Da die parallele Präzession jedoch energetisch geringfügig günstiger als die antiparallele Präzession ist, formieren sich insgesamt etwas mehr Spinmagnete in Richtung von B_0. Dadurch entsteht eine Nettomagnetisierung parallel zu B_0, die einem energetischen Gleichgewichtszustand entspricht.

Die MRT basiert im Wesentlichen auf dem Vorgang, die Spins erst aus diesem Gleichgewicht zu bringen, um dann die Zeit zu messen, in der sie wieder ins Gleichgewicht zurückzukehren. Dies erlaubt Rückschlüsse auf den Wasserstoffgehalt des Gewebes – und damit auf das Gewebe selbst. Bei der Störung der Gleichgewichtsmagnetisierung durch einen hochfrequenten (HF-)Impuls mit einer geeigneten Frequenz greifen die folgenden Mechanismen: Resonanzabsorption und Relaxation.

Resonanzabsorption und Relaxation

Innerhalb des Gleichgewichtszustandes präzessieren die Spins um die Drehachse des äußeren Magnetfeldes. Die Frequenz wird auch als Larmorfrequenz ω bezeichnet. Sie hängt sowohl von einem Proportionalitätsfaktor γ (dem sogenannten gyromagnetischen Verhältnis) des jeweiligen Atomkerns als auch von der Stärke des äußeren Magnetfeldes B_0 ab:

$$\omega = B_0\gamma.$$

Bei den in der MRT verwendeten Feldstärken von über einem Tesla liegt die Larmorfrequenz der ^{1}H-Spins im hochfrequenten Bereich. Trotz konstanter Frequenz präzessieren die Spins nicht synchron um ihre Drehachsen: Stellt man sich die Phase eines Spins wie den Zeiger einer Uhr vor, so liegen die Zeiger zweier Spins tendenziell in völlig unterschiedlichen Zeitzonen. In diesem Zustand ist keine Magnetisierung senkrecht zum äußeren Magnetfeld festzustellen. Man spricht davon, dass die Spins außer Phase liegen.

Zur Störung der Gleichgewichtsmagnetisierung werden HF-Pulse mit der Larmorfrequenz der ^{1}H-Spins (sogenannte Resonanzbedingung) über HF-Spulen senkrecht zum statischen Magnetfeld ausgesendet. Dies ermöglicht, dass die Pulse, die sich wie ein rotierendes Magnetfeld auswirken, kontinuierlich Energie an die Spins der ^{1}H-Kerne abgeben (Resonanzabsorption). Je nach Dauer und Stärke dieses temporären Magnetfeldes, richten sich die Orientierungen der ^{1}H-Spinmagnete neu aus und

präzessieren um die Achse des Pulses, sodass die Nettomagnetisierung um den sogenannten Flipwinkel α entgegen B_0 kippt.

Die resultierende Magnetisierung lässt sich in zwei Komponenten zerlegen: die Längsmagnetisierung in Richtung von B_0 und die zu ihr senkrechte Quermagnetisierung, die um das statische Magnetfeld rotiert und damit eine elektrische Spannung induziert, die als MR-Signal messbar ist.

Nach Anwendung eines 90°-Pulses präzessieren alle Spins um die Achse des Pulses senkrecht zum statischen Magnetfeld. Nach dem Abschalten des Pulses präzessieren sie wieder um das statische Magnetfeld. Dann liegen sie allerdings in Phase, sodass die gesamte ehemalige Längsmagnetisierung in Quermagnetisierung umgewandelt wird.

Die Spins kehren nun jedoch wieder in den vorigen Gleichgewichtszustand zurück (Relaxation). Diesen Vorgang, bei dem sich die Längsmagnetisierung langsam wieder auf- und die Quermagnetisierung schnell abbaut, bezeichnet man als Längs- bzw. Querrelaxation. Die Relaxationsraten hängen zeitlich exponentiell von unterschiedlichen Konstanten ab: Die T1-Relaxationszeit ist die Zeit, in der die Längsmagnetisierung etwa 63 % des vormaligen Gleichgewichtszustands erreicht. Umgekehrt dauert es eine T2-Relaxationszeit, bis die Quermagnetisierung um 63 % zurückgeht. Die genauen Werte der T1- und T2-Konstanten hängen von der Art des Gewebes ab, was maßgeblich zur hohen Gewebekontrastierung in der MRT-Bildgebung beiträgt.

Der Grund für die unterschiedlichen Zeiten liegt darin, dass die T1-Zeit vor allem durch die Gewebeart beeinflusst wird. So ergeben sich durch Magnetfeldinhomogenitäten für Feststoffe oftmals Interaktionen mit Spins anderer Atome. So wirken beispielsweise die Spins von im Gewebe enthaltenem Wasser wie kleine HF-Pulse auf andere Stoffe, wenn sie die entsprechende Larmorfrequenz annehmen. Bei Protonen in Feststoffen spricht man davon, dass sie von einem Gitter umgeben sind, weshalb die T1-Relaxation auch als Spin-Gitter-Relaxation bezeichnet wird.

Die T2-Relaxation beschreibt den Zerfall der Quermagnetisierung durch Wechselwirkungen zwischen eng beieinanderliegenden Spins, die sich gegenseitig außer Phase bringen, die sog. Spin-Spin-Relaxation. In der Praxis fällt das Signal allerdings wesentlich schneller ab als es die Wechselwirkungen erklären. Dies wird als T2*-Relaxation bezeichnet. Die kurzen T2*-Relaxationszeiten entstehen durch die lokalen Magnetfeldschwankungen, die vom Magnetfeld und vom Patienten abhängen. Diese

Schwankungen ändern die Orientierungen der Spins und deren Frequenzen, was jeweils eine zusätzliche Dephasierung bewirkt.

Um die T2-Zeit dennoch zu messen, bedient man sich in der Praxis dem Hilfsmittel des Spin-Echos: Nachdem die Quermagnetisierung erst durch einen 90°-Impuls aufgebaut wurde und sich die Spins in Phase befinden, verschwindet das MR-Signal nach dem Abschalten des Impulses mit der Dephasierung der Spins. Eine halbe Echozeit (TE, engl. echo time) nach dem ersten Impuls sendet man einen 180°-HF-Puls, der bewirkt, dass sich die Ausrichtung der Spins umkehrt. Die Spins durchlaufen nun die lokalen Magnetfeldinhomogenitäten in umgekehrter Reihenfolge, sodass sich ihre Phasen wieder annähern und ein Echo entsteht. Nach TE erreicht das Echo und damit auch das MR-Signal sein Maximum. Während auch das MR-Signal des Echos mit T2* abnimmt, geht das Maximum des Echos mit T2 zurück. Daher kann man mehrere Echos hintereinander erzeugen, um T2 zu berechnen.

Für eine besonders schnelle Bildgebung existiert mit dem Gradienten-Echo eine effizientere Methode zur Erzeugung eines Echos: Hierzu werden nach dem 90°-HF-Puls Magnetfeldgradienten eingeschaltet, die die Dephasierung der Spins beschleunigen. Anstelle eines 180°-Pulses wird nach TE/2 die Polung des Gradienten umgekehrt, sodass die Spins rephasieren und es zu einem Echo kommt, das sein Maximum nach TE annimmt. Wie beim Spin-Echo lässt sich aus der Abnahme der Maxima die T2-Zeit berechnen. Im Unterschied zum Spin-Echo ist die TE beim Gradienten-Echo wesentlich kürzer. Bei der echoplanaren Bildgebung (EPI, engl. echoplanar imaging) werden über bipolare Gradienten innerhalb einer Messung bis zu 128 Echos erzeugt, sodass nur ein HF-Puls benötigt wird, um in unter 100 ms ein Bild aufzunehmen.

Unabhängig von der Erzeugung des Signals, gilt es die räumliche Auflösung der ^{1}H-Atomkerne zu erhalten, indem bei der MRT separat Schichten angeregt werden. Innerhalb der Schichten wird dann eine Frequenz- und Phasenkodierung der Spins vorgenommen, womit sich das Signal entsprechend der Herkunft zerlegen lässt. Dadurch entsteht ein MRT-Bild.

Schichtwahl und Ortskodierung

Um das MR-Signal räumlich aufzuschlüsseln, werden unterschiedliche Mechanismen in x-, y- und z-Richtung angewandt. Die z-Richtung liegt dabei entlang der Richtung des statischen Magnetfeldes. Um eine Schicht auszuwählen, wird zeitgleich zum HF-

Puls ein Gradient entlang der z-Achse erzeugt, sodass sich die Larmorfrequenz der Spins pro Schicht unterscheidet. Die Resonanzbedingung für den HF-Puls erfüllen somit nur Spins innerhalb einer Schicht. Über die Bandbreite des HF-Pulses lässt sich die Dicke dieser Schicht steuern.

Zur Kodierung der x-Koordinate wird während der Messung des Echos ein Gradient in x-Richtung zugeschaltet. Dieser bewirkt, dass die Spins entlang der x-Richtung eine aufsteigende Frequenz aufweisen. Bei dieser Frequenzkodierung lassen sich die einzelnen Frequenzen nach der Messung des Signals mittels einer Fourier-Transformation aus dem Signal herausfiltern.

Zwischen dem HF-Puls und der Messung des Echos wird die Frequenz der Spins kurzzeitig über einen Gradienten in y-Richtung verändert, sodass diese unterschiedlich schnell präzessieren und außer Phase geraten. Diesen Schritt nennt man Phasenkodierung. Mit der Fourier-Transformation lässt sich auch die Phase aus dem Signal herausfiltern. Allerdings gelingt dies nur für eine Phase, weswegen je eine Messung pro Auflösungsschritt in y-Richtung durchgeführt werden muss. Das beeinflusst die Dauer des MR-Experiments maßgeblich. Da die TE-Zeit wesentlich kürzer ist als die Wiederholungszeit des HF Pulses, kann die Messung weiterer Schichten in der Zeit bis zum nächsten HF-Puls vorgenommen werden. Mit dieser sogenannten Mehrschichtsequenz können dreidimensionale Daten generiert werden.

Die Ergebnisse der drei Kodierungsschritte werden für jede Messung in eine Rohdatenmatrix geschrieben, die als K-Raum bezeichnet wird. Über eine zweidimensionale Fourier-Transformation können aus den Rohdaten im K-Raum nun die zugehörigen Signale aus der MRT-Messung allen x- und y-Koordinaten zugeordnet werden, wodurch ein MRT-Bild entsteht.

Eine vollständige Pulssequenz besteht aus HF-Puls, der Schicht- bzw. Ortskodierung und dem Auslesen des Echos. Die Zeit für die Messung einer Phase entspricht der Wiederholungszeit (TR, engl. repetition time) des HF-Pulses. Das Signal und die Grauwerte des MRT-Bildes hängen von der gewählten Gewichtung ab. So kann über TR und TE gesteuert werden, wie das Signal gewichtet ist: nach der Protonendichte (lange TR und kurze TE), der T1-Zeit (kurze TR und kurze TE) oder der T2-Zeit (lange TR und lange TE).

Auf diesen Techniken basieren die MRT-Sequenzen, die im nächsten Abschnitt beschrieben sind. Sie kamen bei den Schlaganfallpatienten aus der I-KNOW-Studie zum Einsatz. Ihre Bilddaten sind auch in die Modellierung in dieser Arbeit eingeflossen.

2.4.2 MRT-Sequenzen zur Auswahl der Therapieverfahren

Fluid-attenuated Inversion Recovery

Anhand der FLAIR-Sequenz kann eine Gehirnblutung beim Schlaganfall sicher ausgeschlossen werden [Vert et al., 2017]. Die Besonderheit dieser Spin-Echo-Sequenz besteht darin, dass ihr ein 180°-Puls vorgeschaltet ist, wodurch das Signal von freier Gewebeflüssigkeit (CSF, engl. cerebrospinal fluid) unterdrückt wird. So erlaubt die FLAIR- gegenüber der T2-Sequenz eine bessere Detektion von ischämischen Läsionen und anderen Pathologien, die sonst von freier Gewebeflüssigkeit überlagert sind, wie u. a. in der Nähe der Ventrikel [Billebaut und Wameling, 2012; Vert et al., 2017]. Nach der sogenannten Inversionszeit des CSFs von etwa 2,5 Sekunden ist dessen Längsmagnetisierung genau null. In diesem Moment wird der 90°-HF-Puls angewandt, sodass keine Resonanzabsorption des Gewebewassers stattfindet und dieses bei der FLAIR-Sequenz keinen Anteil am Signal hat [Billebaut und Wameling, 2012]. Damit gilt die FLAIR als Goldstandard zur Segmentierung von irreversiblen Läsionen [Scalzo et al., 2012].

Da das Erkennen von ischämischen Läsionen mittels FLAIR in mehr als 50% der Fälle erst nach über drei Stunden möglich ist, eignet sie sich jedoch nicht zur frühen

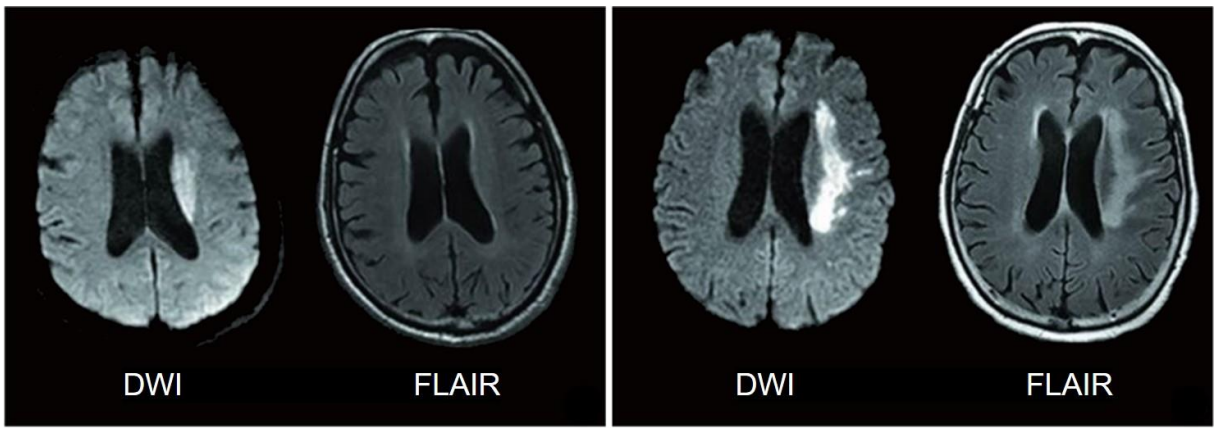

Abbildung 2: DWI-FLAIR-Mismatch: In der linken Aufnahme ist eine Läsion in der DWI bereits sichtbar, jedoch noch nicht in der FLAIR, d. h. ein DWI-FLAIR-Mismatch ist vorhanden. Dies weist auf ein hyperakutes Infarktgeschehen hin, was als Indikator gilt, dass das Zeitfenster für eine erfolgreiche Thrombolyse noch nicht abgelaufen ist. Rechts hingegen ist ein Infarkt sowohl in der DWI als auch in der FLAIR sichtbar (Quelle: B. J. Kim et al., 2014).

Infarkterkennung [Yilmaz, 2015]. Stattdessen ist dann die diffusionsgewichtete Bildgebung zu bevorzugen. Die gemeinsame Betrachtung der beiden Sequenzen liefert mit dem DWI-FLAIR-Mismatch (siehe Abbildung 2) einen Indikator, ob es sich um einen sehr frühen (keine FLAIR-Läsion) oder schon mehrere Stunden alten Infarkt handelt, sodass geeignete Therapien für Patienten mit unklarem Zeitfenster ausgewählt werden können [Yilmaz, 2015; Vert et al., 2017].

Diffusionsgewichtete MRT

Die diffusionsgewichtete Bildgebung bietet die früheste Möglichkeit, die pathophysiologischen Veränderungen des Schlaganfallgewebes zu erkennen [Yilmaz, 2015]. Laut gängiger Theorie führt der ischämiebedingte Sauerstoffmangel (Hypoxie) schon frühzeitig zur bereits beschriebenen Störung der Natrium-Kalium-Pumpe, wodurch vermehrt Gewebewasser ins Zellinnere gelangt (zytotoxisches Ödem) [Neil, 1997; Brown und Semelka, 2003]. Innerhalb der Zellen ist die Bewegung der Wasserstoffmoleküle durch den Zellradius stark eingeschränkt. Verstärkt wird dieser Effekt noch durch eine erhöhte Viskosität im Zellinneren, ausgelöst durch den energiebedingten Zerfall des Zytoskeletts und der Zellorganellen. Das erhöhte Volumen der angeschwollenen Zellen führt zusätzlich zu einer eingeschränkten Bewegung der Wasserstoffmoleküle außerhalb der Zellen [Yilmaz, 2015]. Mit der MRT sind diese Molekülbewegungen messbar, die auf leichten Konzentrationsunterschieden in zwei Umgebungen basieren und auch als *Diffusion* bezeichnet werden [Brown und Semelka, 2003].

Üblich ist die Messung der Diffusion mithilfe einer EPI-Sequenz, wobei die Spin-Echo-Sequenz wie beschrieben leicht modifiziert wird. So werden während der Dephasierungs- und Rephasierungsphasen starke Gradienten-Impulse erzeugt, die die Dephasierung der Spins verstärken. Die Idee dahinter ist, dass Spins bzw. ^{1}H-Atomkerne, die sich zwischen den zwei Impulsen bewegt haben, ungleiche Effekte durch die Impulse erfahren und daher nicht rephasieren. Damit wird das Signal geschwächt. Der Diffusionskoeffizient D lässt sich aus der Gleichung für das (geschwächte) Signal

$$S = S_0 \cdot \exp(-b \cdot D)$$

berechnen, wobei das Ausgangssignal $S_0 = e^{-TE/\mathrm{T2}}$ entspricht und der *b-Wert* die Sensitivität des Signals für Bewegung festlegt. Dieser kann über die Gradientenfeld-

stärke G, den zeitlichen Abstand der Impulse T und die Dauer der Diffusionsgewichtung t gesteuert werden:

$$b = \gamma^2 \mathrm{G}^2 \mathrm{t}^2 \left(T - \frac{t}{3}\right).$$

Im Gehirn wird die Messung des Diffusionskoeffizienten vor allem durch den Blutfluss in winzig kleinen, zufällig verlaufenden Gefäßen innerhalb eines Voxels beeinflusst. Daraus ergeben sich Signalverluste, die nicht von einer Diffusion unterschieden werden können, weswegen man bei In-vivo-Messungen von einem *scheinbaren Diffusionskoeffizienten* spricht (ADC, engl. apparent diffusion coefficient). Von der Diffusion sind ebenso kaum Effekte durch T2-Relaxation zu differenzieren, die bei Spin-Echo-Sequenzen auftreten und bei DWI-Sequenzen durch relativ lange Echozeiten noch verstärkt werden. Deshalb berechnet man Karten für verschiedene b-Werte (üblicherweise zwischen 0 und 1000), aus denen eine ADC-Karte extrapoliert wird. Diffusion ist im menschlichen Körper nicht gleichmäßig, sondern richtungsabhängig, weshalb der ADC nur eine vereinfachte Beschreibung der Diffusion repräsentiert [Brown und Semelka, 2003].

Im nächsten Abschnitt ist als weiteres MR-Verfahren die Time-of-Flight-MR-Angiographie beschrieben, die eine Visualisierung des Blutflusses innerhalb der Gefäße sowie der Unterbrechungen aufgrund pathologischer Zustände wie Stenosen oder Verschlüssen ermöglicht. Dies gibt vor allem Aufschluss über die Lokalisation eines Gerinnsels und bildet die Indikationsbasis für eine zielgerichtete Rekanalisation durch Thrombektomie.

Time-of-Flight-MR-Angiographie

Mit der MR-Angiographie lassen sich Gefäße anhand des laminaren Blutflusses visualisieren, sodass die Durchlässigkeit der Gefäße evaluiert werden kann und Unterbrechungen des Blutflusses durch Stenosen oder Gerinnsel erkennbar werden. Grundsätzlich unterscheidet man bei der MR-Angiographie zwischen Bright- und Dark-Blood-Ansätzen, wobei die Time-of-Flight-Angiographie zur ersten Gruppe gehört, weil sie Blut hell darstellt [Brown und Semelka, 2003]. Die schnellste MR-Angiographie-Methode wurde ebenfalls bei den I-KNOW-Patienten angewandt. Eine Time-of-Flight-Angiographie zur Visualisierung eines proximalen Verschlusses der mittleren Gehirnschlagader mit und ohne Kontrastmittel ist in Abbildung 3 dargestellt.

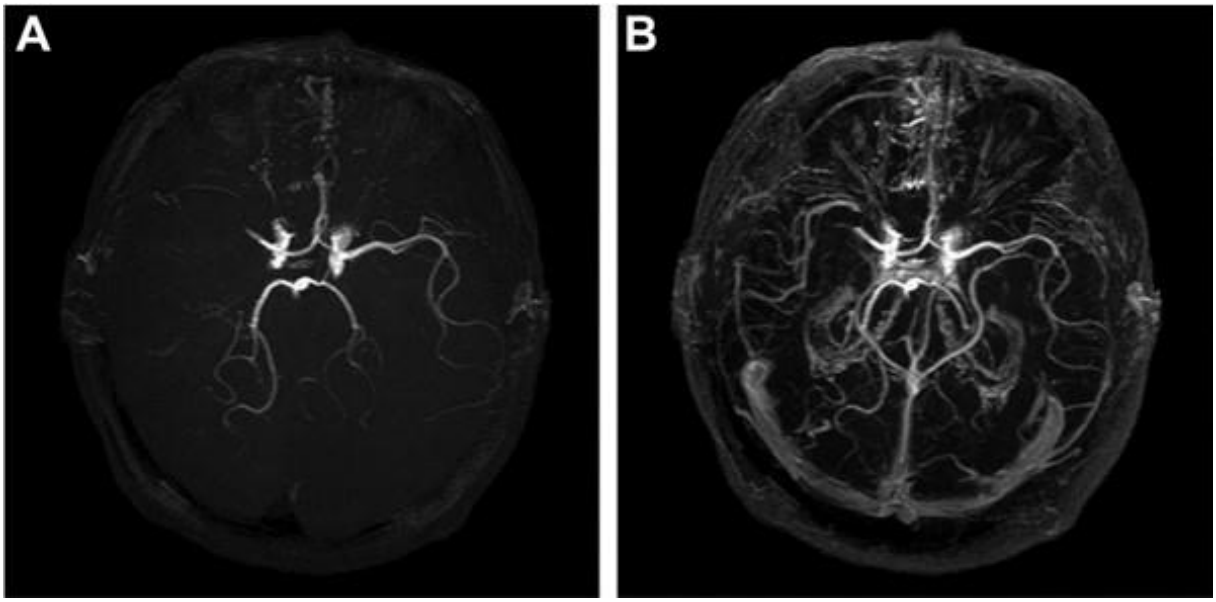

Abbildung 3: Time-of-Flight-Angiographie für einen proximalen rechtsseitigen Verschluss der mittleren Gehirnschlagader vor (A) und nach (B) Kontrastmittelgabe. Der Verschluss lässt sich in beiden Darstellungen anhand des scheinbar asymmetrischen Gefäßverlaufs erkennen. Hierbei gilt es zu beachten, dass sich die rechte Gehirnhälfte in der neuroradiologischen Ansicht üblicherweise auf der jeweils linken Seite der Darstellungen befindet (Quelle: Sohn et al., 2007).

Bei der Time-of-Flight-MRA wird das Gewebe der zu messenden Schicht zunächst durch eine Gradienten-Echo-Sequenz mit kurzen TRs und moderatem Flipwinkel gesättigt, sodass in die Schicht einströmendes Blut den größten Anteil am MR-Signal ausmacht. Diese Technik hat gegenüber röntgenbasierten Verfahren u. a. den Vorteil, dass Messungen mehrmals wiederholt werden können, weil kein Kontrastmittel benötigt wird. Da die Richtung des Blutflusses bei der Time-of-Flight-MRA normalerweise außer Acht gelassen wird, genügt bei ihr eine Messung. Zur Unterscheidung eng aneinanderliegender Arterien und Venen mit verschiedenen Blutflussrichtungen lässt sich das Signal des Blutes in einer benachbarten Schicht durch einen Vorsättigungsimpuls schwächen, wodurch die Richtung des Blutflusses festgestellt werden kann [Brown und Semelka, 2003].

Durch Mehrschicht-Sequenzen können neben zweidimensionalen Schichten auch dreidimensionale Aufnahmen erstellt werden. Die zu untersuchenden Regionen sollten allgemein so ausgewählt werden, dass die Gefäße nicht innerhalb der Schichten verlaufen, da der Blutfluss dort schnell gesättigt wird. Wenn die Flussrichtung senkrecht zur Schicht steht und die Schicht hinreichend dünn ist, bleibt das Signal über die gesamte Schichtdicke vom Gewebesignal unterscheidbar [Brown und Semelka, 2003].

Das Problem der schnellen Sättigung des Blutes tritt insbesondere bei dreidimensionalen MR-Angiographie-Verfahren auf, weshalb sich diese in erster Linie für die Visualisierung größerer Gefäße mit entsprechend schnellem Blutfluss eignen. Eine Möglichkeit zum Ausgleich der Sättigung des Blutes besteht in der Variation des Flipwinkels. Eintrittsseitig zur Schicht kann ein kleiner Flipwinkel für eine geringe Blutsättigung gewählt werden, wohingegen sich austrittsseitig ein großer Flipwinkel für eine hohe Sättigung anbietet [Brown und Semelka, 2003].

Perfusionsgewichtete MRT

Zur Bestätigung der Rekanalisationsindikation nach Ablauf des 4,5 Stunden-Zeitfensters (oder bei unklarem Zeitfenster) werden Informationen über den kapillaren Blutfluss benötigt [Ringleb et al., 2021]. Dies betrifft immerhin knapp 75 % der Schlaganfallpatienten [Tong et al., 2012]. Ob eine hinreichend große Penumbra vorliegt, lässt sich anhand der Perfusionsbildgebung klären [Zaro-Weber et al., 2016; Ringleb et al., 2021].

Insbesondere nach Ablauf des 4,5-Stunden Zeitfensters sollte für Patienten mit klinischer Rekanalisationsindikation deshalb zusätzlich eine Perfusionsbildgebung erfolgen [Ringleb et al., 2021]. Dazu wird ein Gadolinium-haltiges Kontrastmittel injiziert, welches auch als Bolus bezeichnet wird, wobei dessen initiale Passage durch das Gefäßbett mittels einer EPI-Sequenz anhand wiederholter Aufnahmen in kurzen Abständen verfolgt wird.[15] Aus den vierdimensionalen Aufnahmen lassen sich zeitabhängige Konzentrationskurven des Kontrastmittels für jedes Voxel des Gehirns bestimmen. Diese geben Aufschluss darüber, wann wie viel Kontrastmittel im Gewebe ankommt (siehe Abbildung 4). Im Rahmen verschiedener kinetischer Modelle, die sich danach unterteilen lassen, ob die Konzentrationskurven des Gewebes um diejenigen für das arterielle Zuflussgefäß (AIF, engl. arterial input function) korrigiert werden,[16] können die wichtigsten Perfusionsparameter aus den Konzentrationszeitkurven abgeleitet werden [Sourbron, 2010; Heit und Wintermark, 2016].

Das zerebrale Blutvolumen (CBV, engl. cerebral blood volume) ist die Blutmenge innerhalb eines Voxels und entspricht der Fläche unter der Kontrastmittelkurve. Es

[15] Alternativ zu einem Kontrastmittel kann die PWI-Bildgebung auch mittels Arterial Spin Labelling durchgeführt werden [Sourbron, 2010].

[16] Dies ist u. a. relevant um die Menge des individuell verabreichten Kontrastmittels zu berücksichtigen, sodass PWI-Parameter zwischen verschiedenen Patienten vergleichbar bleiben [Forkert et al., 2014].

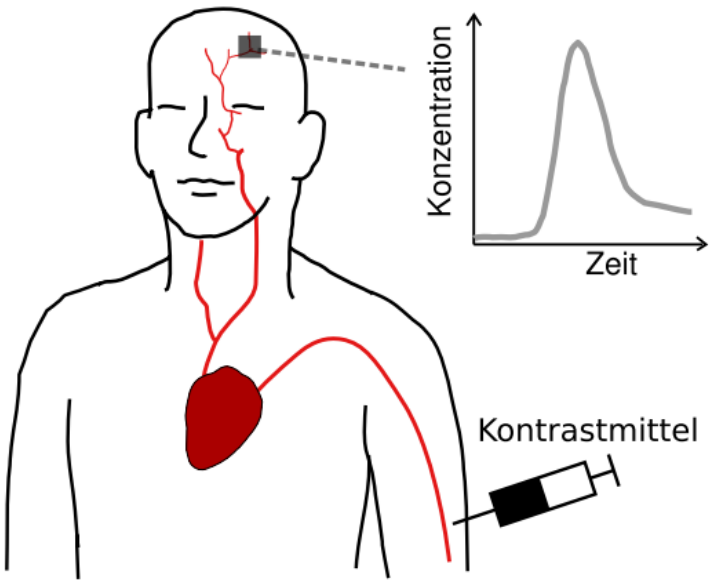

Abbildung 4: Messprinzip der perfusionsgewichteten Bildgebung: Nach Injektion der Kontrastmittelgabe werden Aufnahmen des Gehirngewebes in kurzen Abständen durchgeführt. So kann die Konzentration des Kontrastmittels über die Zeit – wie in dieser Skizze schematisch dargestellt – für jedes Voxel des Gehirngewebes nachvollzogen werden (Quelle: Kellner, 2014).

wird in ml/100 g angegeben. Die Zeit, die das Blut durchschnittlich für eine Passage durch ein Voxel benötigt, ist als mittlere Transitzeit definiert (MTT, engl. mean transit time). Der zerebrale Blutfluss (CBF, engl. cerebral blood flow) entspricht der Menge des Bluts, das innerhalb einer vorgegebenen Zeit durch ein Voxel fließt. Diese Menge wird in ml/100 g/min angegeben. Nach dem sogenannten „central volume principle“ besteht zwischen diesen Parametern der Zusammenhang CBF = CBV/MTT.

Zusätzlich wird bei der Perfusionsbildgebung zumeist noch die Zeit bis zur maximalen Kontrastmittelanreicherung in einem Voxel (TTP, engl. time to peak) und die Zeit bis zum Maximum der Residuenfunktion (T_{max}, engl. time-to-maximum of the residue function) bestimmt [Sotoudeh et al., 2019]. Letztere spiegelt in erster Linie die Verzögerung des Kontrastmittels zwischen dem Ort, an dem die AIF (engl. arterial input function) gemessen wurde, und dem Gewebe wider [Wouters et al., 2017].

Die Kontrastmittelkonzentration $C(t)$ lässt sich als Faltung der Residuenfunktion des Kontrastmittels $R(t)$ mit der AIF ausdrücken, $C(t) = CBF \cdot R(t) * AIF(t)$. Die Residuenfunktion lässt sich dann über die Entfaltung berechnen, wobei der CBF direkt über das Maximum der Residuenfunktion ($R(t = 0) = 1$) ablesbar ist [Sourbron, 2010]. Die mittlere Transitzeit ist in diesem Ansatz als das Integral über $R(t)$ definiert. Als problematisch erweist sich die Bestimmung einer geeigneten AIF bei diesem dekon-

volutionsbasierten Ansatz, da u. a. die Auflösung der bildgebenden Verfahren in der Regel nicht ausreicht, um die Konzentration ausschließlich innerhalb eines einzigen, oft winzigen Gefäßes zu messen (sogenannter *partial volume effect*). In der klinischen Praxis werden deshalb bei diesem Ansatz meist AIFs der vorgelagerten Arterien ausgewählt [Sourbron, 2010].

Ein niedriger CBF oder eine hohe Zeitverzögerung des Kontrastmittels liefern Hinweise für eine Durchblutungsstörung, wobei diese Marker nichts darüber aussagen, wie viel Blut tatsächlich im Gewebe ankommt. Weist das Gewebe zugleich einen hohen CBV auf, so ist dieses Mismatch ein Anzeichen für eine Penumbra [Heit und Wintermark, 2016]. Deren Ausmaß dient regelmäßig der Patientenselektion für die rekanalisierenden Therapien, für die normalerweise eine relative und absolute Mindestgröße von 20 % des Infarktkerns bzw. 10 ml für das Mismatch vorausgesetzt wird.

Eine rein visuelle Betrachtung des Mismatch führt jedoch häufig zu einer Überschätzung des tatsächlichen Läsionsoutcomes [Campbell und Parsons, 2018]. Für die Differenzierung von Kern und Penumbra verwenden Softwareanbieter und Institute in der Praxis eine große Bandbreite an Methoden und Schwellwerten [Heit und Wintermark, 2016]. Nach neuestem Stand scheint $T_{max} > 6$ s ein angemessener Schwellwert zur Abgrenzung des Durchblutungsdefizits zu sein, wonach sich ischämisches Gewebe vom gesundem Gewebe abgrenzen lässt [Campbell und Parsons, 2018].[17] Der Infarktkern lässt sich bei der MRT aus der diffusionsgewichteten Bildgebung bestimmen. Ein ADC < 620 mm^2/s gilt dabei als ein kritischer Schwellwert [Purushotham et al., 2013].

In Abbildung 5 ist die Bildung des PWI-DWI-Mismatches am Beispiel einer akuten Schlaganfallpatientin aus der I-KNOW-Studie visualisiert. Hier lässt sich sofort eine Durchblutungsstörung in der rechten Gehirnhälfte anhand der T_{max}- und MTT-Karten identifizieren. Der Bereich mit relativ niedrigem CBF bestätigt sich in der ADC-Karte als DWI-Läsion bzw. Infarktkern. In einer PWI-DWI-Mismatch-Karte werden der durchblutungsgestörte Bereich ($T_{max} > 6$ s) und der Infarktkern (ADC < 620 mm^2/s) schwellwertbasiert voneinander abgegrenzt. Trotz frühzeitiger Thrombolyse-Behandlung wuchs der Infarkt weiter stark an, sodass auf der Follow-up-FLAIR eine Infarzierung über nahezu das gesamte Durchblutungsdefizit sichtbar wird.

[17] Analog zur MRT ist das Mismatch-Konzept auch auf die CT übertragbar. Da dort keine Diffusion bestimmt werden kann, wird für die CT der Kern auf Basis der CT-PWI ermittelt. Als am häufigsten genutzte Schwellwerte zur Bestimmung des Infarktkerns gelten bei der CT (1) ein CBV < 2 ml/100 g und (2) ein um 38–50 % reduzierter CBF gegenüber der kontralateralen Hemisphäre [Heit und Wintermark, 2016].

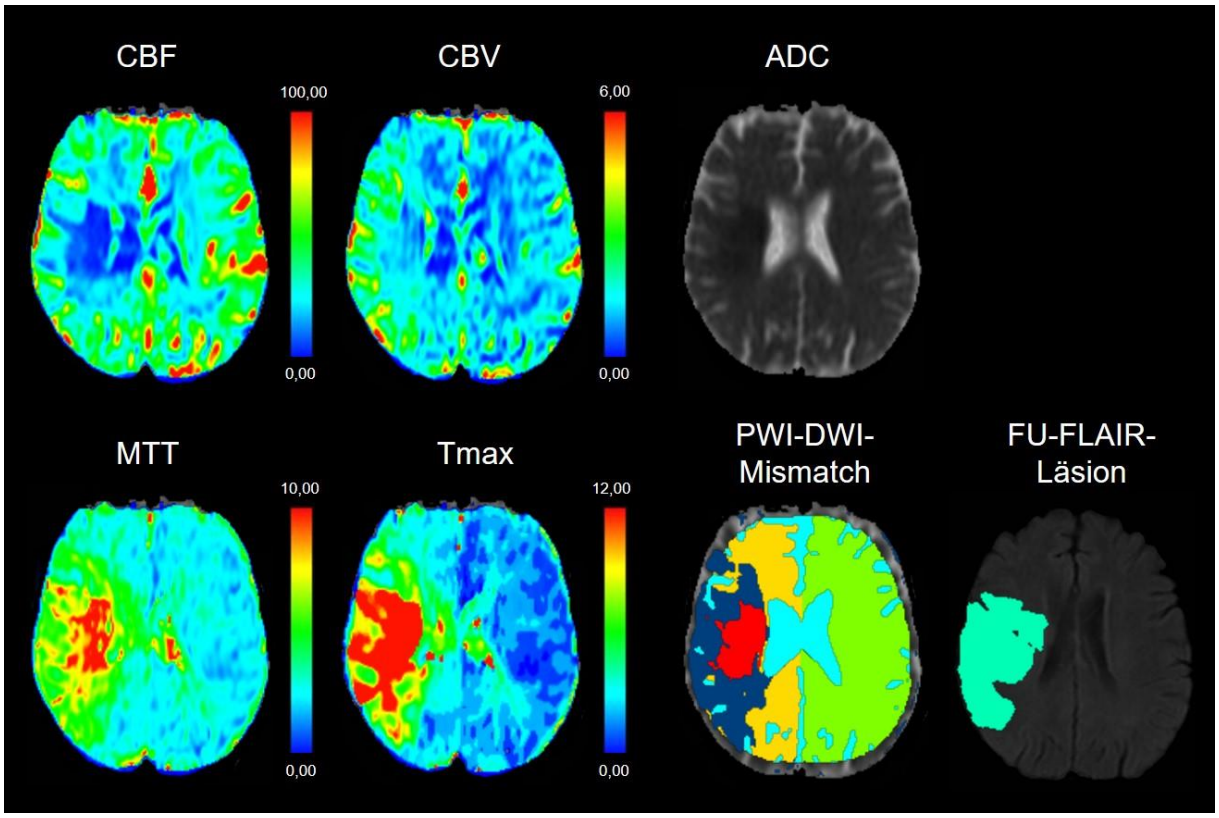

Abbildung 5: PWI-DWI-Mismatch für eine ausgewählte Patientin (69 Jahre, NIHSS = 10): Die zeitabhängigen PWI-Karten MTT und T_{max} zeigen eine Durchblutungsstörung für die rechte Gehirnhälfte an. Der relativ niedrige CBF im Vergleich zur linken Gehirnhälfte und die ADC-Läsion (dunkler Bereich neben dem Ventrikel) definieren das Gebiet des Infarktkerns. Eine Bestimmung der Penumbra wird hier schwellwertbasiert mittels PWI-DWI-Mismatch getroffen (T_{max} > 6 s, ADC < 620 mm^2/s). Der finale Infarkt hat sich deutlich über den Infarktkern hinaus auf nahezu das gesamte Mismatch ausgeweitet (Quelle: I-KNOW-Daten).

Kritik am Mismatch-Konzept

Trotz der Möglichkeit, Patienten auch nach den kritischen Zeitfenstern für Rekanalisationstherapien auszuwählen, zeigt das Mismatch-Konzept noch einige Schwächen hinsichtlich der Definition der Penumbra. So lässt sich der Zustand und die Entwicklung des infarzierten Gewebes (Kern) nicht abschließend anhand einer einzelnen Messung während der Akutphase bewerten. Der Verlauf ist stark zeitabhängig, denn Infarkte verlaufen dynamisch und können unterschiedlich schnell voranschreiten. Häufig kommt es bei einer frühen Rekanalisation vor, dass Teilgebiete oder in manchen Fällen sogar der gesamte als Infarktkern spezifizierte Bereich noch überleben (Pseudonormalisierung). Zudem reagieren verschiedene Gewebe- und Zelltypen unterschiedlich sensibel auf eine Ischämie. Beispielsweise benötigt graue Substanz wesentlich mehr Energie als weiße Substanz und ist daher besonders stark vom Zelltod bedroht.

Deshalb erweist sich ein Schwellwert auch bezüglich des Perfusionsdefizits als ungeeignet. Teilweise überlebt noch ein Teil des minderperfundierten Gebiets trotz ausbleibender Reperfusion („benigne Oligämie“). Darüber hinaus variiert die Resistenz des Gewebes gegenüber Ischämien mit ihrer Dauer und Tiefe sowie von Patient zu Patient z. B. aufgrund des Alters, systemischer Vorerkrankungen wie Diabetes mellitus oder Gehirnerkrankungen wie Leukoaraiose etc. [McKinley et al., 2016; Goyal et al., 2020].

Selbst wenn nur für einen kleinen Teil der eingelieferten Fälle eine genauere Patientenselektion für die geeigneten Therapiemaßnahmen vorgenommen werden kann, hat dies aufgrund der hohen Prävalenz und Einzelfallkosten immer noch einen erheblichen Effekt auf die Gesundheit der zu behandelnden Personen und die Kosten im Gesundheitssystem. Anstelle von einfachen Schwellwerten zur Abgrenzung von Infarktkern und Penumbra, wie beim PWI-DWI-Mismatch, erlauben binäre Klassifikationsmodelle die gleichzeitige Berücksichtigung aller vorhandenen Informationen aus der akuten Bildgebung und den klinischen Informationen für die Vorhersage des geschädigten Gewebes, das pro Voxel der Klasse Läsion oder Nichtläsion in der Follow-up-Untersuchung entspricht. Aus diesem Grund werden zunehmend binäre Klassifikationsmodelle erforscht und in der akuten Schlaganfalldiagnostik eingesetzt. So lässt sich das Gewebeoutcome beim Schlaganfall genauer vorhersagen, was eine bessere Patientenselektion für die Rekanalisationstherapien ermöglicht.

2.5 Binäre Klassifikationsmodelle zur Vorhersage des Gewebeschicksals für eine personalisierte Therapieentscheidung

Seit mehr als 20 Jahren werden in der Schlaganfalldiagnostik binäre Klassifikationsmodelle zur Vorhersage des Gewebeoutcomes erprobt [Schlaug et al., 1999].[18] Sie gehören zu den Verfahren des überwachten maschinellen Lernens (engl. machine learning), welches wiederum ein Teilgebiet der künstlichen Intelligenz ist.

Mit dem überwachten Machine Learning wird angestrebt, die Parameter einer Modellgleichung so zu optimieren, dass die Gleichung für deren Input möglichst genau den zugehörigen Output bestimmt. Dies geschieht auf Basis von komplementären Datensätzen, die jeweils aus Input und Output bestehen. Genauigkeit ist in diesem Zusammenhang über eine sogenannte Zielfunktion definiert, die den Output des Modells

[18] Genau genommen ist auch das PWI-DWI-Mismatch in Form eines Entscheidungsbaums ein binäres Klassifikationsmodell.

mit dem tatsächlichen Output abgleicht. Die Optimierung basiert auf einem Lernalgorithmus, der das Modell sukzessive anhand des Feedbacks der Zielfunktion verbessert.

Dieser Prozess ist als Training des Modells zu verstehen, in welchem die verwendeten Wertepaare die entsprechenden Trainingsdaten bilden. Im Gegensatz zu Regressionsmodellen kann der Output bei Klassifikationsmodellen nur diskrete Werte annehmen und bei binären Klassifikationsproblemen bestehen für den Output genau zwei Klassen. Eine ausführliche Beschreibung von einer Vielzahl an Methoden des überwachten Machine Learnings findet sich bei [Hastie et al., 2009].

Die meisten Klassifikationsalgorithmen geben nicht direkt eine Klasse, sondern zuerst einen Score zwischen null und eins pro Klasse aus. Dabei gilt: Je höher der Score, desto stärker vermutet das Modell, dass der Output der jeweiligen Klasse entspricht. Für binäre Klassifikationsmodelle entspricht die Summe der Scores für beide Klassen in der Regel dem Wert eins, weshalb die Modelle nur einen Score für eine der Klassen zurückgeben. Nach Konvention ist dies die seltenere der beiden Klassen.

Im Allgemeinen werden die Trainingsdatenpaare beim Machine Learning auch als Features (Input) und Target (Output) bezeichnet. Bei der Vorhersage des Gewebeschicksals werden diese Datenpaare auf Voxelebene erhoben. Als Features werden üblicherweise die beschriebenen PWI- und DWI-Parameter aus der akuten Bildgebung verwendet. Das Target ist das binäre Gewebeoutcome (kurz Outcome), das aus einer Verlaufsuntersuchung (FU, engl. follow-up) mit einer FLAIR-Sequenz erhoben wird.

Durch die Anwendung eines binären Klassifikationsmodells auf einen neuen Patientendatensatz ergibt sich eine Heatmap des Gehirns, die anzeigt, welche Hirnregionen vom Schlaganfall am wahrscheinlichsten geschädigt werden. Darüber hinaus lässt sich jedes Voxel dieser Karte aus kontinuierlichen Werten zwischen null und eins über eine schwellwertbasierte Binarisierung in die Klassen Läsion und Nichtläsion einteilen. Die Ergebnisse solcher Modelle werden immer häufiger zur Einschätzung des Krankheitsverlaufs in der klinischen Praxis hinzugezogen, weil sie Prognosen über die Wahrscheinlichkeit weiterer Gewebeschäden bieten. Analog zum Mismatch unterstützen sie die Beurteilung, ob ein Patient noch von einer erfolgreichen Rekanalisationstherapie profitieren würde [Mainali et al., 2021].

Eine grundsätzliche Voraussetzung für das Trainieren eines robusten Machine-Learning-Modells ist eine heterogene Datenbasis. Sie sollte möglichst das gesamte

Spektrum an realistisch auftretenden Fällen abdecken, sodass das Modell auf ausreichend Evidenz basiert, um es in Notfallsituationen auf neue akute Patientendatensätze zu übertragen.

Beim Training stellt sich zusätzlich die Frage nach dem Trade-off zwischen der Komplexität und Verallgemeinerbarkeit eines Modells. Es ist möglich, eine ausreichend komplexe Modellfunktion zu trainieren, die alle Datensätze in den Trainingsdaten zumindest beinahe fehlerfrei vorhersagt. Bei einem solch niedrigen Bias scheint es, als bestünden aussagekräftige Zusammenhänge zwischen den Features und dem Target, die vom Modell erkannt und in realistische Klassifikationen umgewandelt werden. In der Praxis stellt sich für solche Modelle allerdings häufig heraus, dass sie zu spezifisch an die einzelnen Trainingsdatensätze angelehnt sind und anstelle allgemeiner Trends für jedes Trainingsdatenpaar eine eigene Regel erlernt haben. Dies führt bei der Anwendung für ungesehene Datensätze mit leicht unterschiedlichen Voraussetzungen zu einer starken Streuung (hohen Varianz) in den Prognosefehlern. Daher spricht man allgemein von einer Überanpassung an die Trainingsdaten, deren zugrundeliegende Problematik auch als Bias-Varianz-Dilemma (engl. bias-variance tradeoff) bezeichnet wird.

Zur Steuerung des Trade-offs zwischen Genauigkeit und Stabilität werden zwei modellspezifische Strategien eingesetzt. Mit sogenannten Regularisierungstermen in den Zielfunktionen vieler Lernalgorithmen versucht man zum einen komplexere Lernmuster zu bestrafen. Im Zweifel werden so weniger und einfachere Mechanismen erlernt, die eher für eine große Menge der Trainingsdaten zutreffen als für wenige einzelne Trainingsdatenpaare. Zum anderen hängen viele Modelle von sogenannten Hyperparametern ab, wie z. B. der Tiefe eines Entscheidungsbaumes, über die sich verschiedene modellagnostische Aspekte steuern lassen. Diese werden nicht direkt aus den Daten erlernt, sondern sind vorgegeben und müssen im Sinne des Bias-Varianz-Trade-offs sorgfältig kalibriert werden. Daher werden sie auch als Tuning-Parameter bezeichnet.

Um während der Entwicklung zu testen, wie gut ein Modell verallgemeinerbar ist, wird daher üblicherweise ein Teil der Trainingsdaten vom Training exkludiert und als Testdatensatz zur Modellvalidierung zurückbehalten. Um eine willkürliche Einteilung in Trainings- und Testdatensätze zu vermeiden, wird dieser Prozess häufig im Rahmen einer Kreuzvalidierung systematisch für verschiedene Kombinationen aus Trainings- und Testdaten wiederholt durchgeführt. Der genaue Prozess bei der Vorhersage des

Gewebeoutcomes und die wichtigsten Metriken zur Validierung von binären Klassifikationsmodellen beim ischämischen Schlaganfall werden im nächsten Abschnitt beschrieben.

2.5.1 Validierung und Metriken

Leave-One-Patient-Out-Kreuzvalidierung

Eine Besonderheit der voxelbasierten Prädiktion des Gewebeoutcomes ist, dass sich die Prognose für einen Patienten aus vielen einzelnen Vorhersagen auf Voxelebene zusammensetzt. Daher muss bei der Entwicklung eines Vorhersagemodells berücksichtigt werden, dass die Voxel eines Patienten stets entweder zum Training oder zum Testen verwendet werden; nicht jedoch für beides.

Ein übliches Schema für die Validierung dieser Modelle besteht darin, die Voxel jedes Patienten genau einmal als Testdatensatz zurückzubehalten. Alle jeweils anderen Patientendatensätze werden zum Modelltraining verwendet. So lässt sich ein Modell an allen Patienten testen und dabei auf der maximalen Anzahl an Patienten trainieren. Im Rahmen dieser Arbeit wird dieses Schema als *Leave-One-Patient-Out-Kreuzvalidierung* (LOPO) bezeichnet.

Für jede Vorhersage eines Patientendatensatzes lässt sich deren Prognosequalität anhand verschiedener Metriken evaluieren. Einige Metriken hängen davon ab, welcher Schwellwert für die Binarisierung des Prognosescores verwendet wird.

Schwellwert

Zur Prognose des Gewebeoutcomes berechnen binäre Klassifikationsmodelle in der Regel einen Prognosescore zwischen null (Nichtläsion) und eins (Läsion). Dabei gilt: Je höher der Prognosescore, desto höher ist die Wahrscheinlichkeit, dass das Voxel untergeht (Läsion) und umgekehrt. Die finale Einteilung der Voxelvorhersagen in die unterschiedlichen Outcomeklassen (Binarisierung) wird schwellwertbasiert getroffen. Voxel, deren Prognosescore größer oder gleich jenem Schwellwert ist, werden als Läsionsvoxel vorhergesagt. Voxel, deren Prognosescore kleiner als der Schwellwert ist, werden als Nichtläsionsvoxel eingestuft.

Confusion Matrix

Anhand des bekannten Voxeloutcomes lassen sich die Wahrheitswerte der binarisierten Prognosen (wahr oder falsch) in einer sogenannten Confusion Matrix, im Nachhinein in vier Möglichkeiten unterteilen: true-positive (TP), false-positive (FP), true-negative (TN) und false-negative (FN). Aus den Häufigkeiten dieser vier Fälle, die im Folgenden mit Betragsstrichen gekennzeichnet sind, lassen sich u. a. die folgenden schwellwertabhängigen Metriken definieren. Diese werden häufig zur Validierung von binären Klassifikationsmodellen verwendet.

Accuracy

Die Accuracy misst, wie häufig ein Modell im Durchschnitt mit seinen Prognosen richtig liegt. Sie ist definiert als:

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

Allein ist die Metrik Accuracy bei unausgewogenen binären Klassifikationsproblemen allerdings wenig aussagekräftig, da sie nicht zwischen unterschiedlichen Outcomes differenziert, sondern richtige Vorhersagen beider Klassen gleich stark gewichtet. So würde z. B. die Accuracy eines Modells, welches alle Voxel als gesund vorhersagt, allein aufgrund der vergleichsweise geringen Infarktvolumina einen sehr hohen Wert aufweisen. Für die meisten Klassifikationsprobleme, insbesondere für die Vorhersage des Gewebeoutcomes, ist jedoch die korrekte Vorhersage des selteneren Outcomes relevanter, weil der häufigere Fall (gesundes Gewebe) recht schnell und einfach im Vergleich zu den geschädigten Läsionsgebieten zu erkennen ist. Daher wird die Accuracy selten als Zielmetrik zur Beurteilung von binären Klassifikationsmodellen herangezogen.

Dice-Koeffizient

Der Dice-Koeffizient ist definiert als:

$$Dice\text{-}Koeffizient = \frac{2|TP|}{2|TP| + |FP| + |FN|}$$

Seine Interpretation ist besser ableitbar aus der mengentheoretischen Definition:

$$Dice\text{-}Koeffizient = \frac{2|X \cap Y|}{|X| + |Y|}$$

Bei dieser Definition ist X die Menge der Läsionsvoxel und Y die Menge der prognostizierten Läsionsvoxel. Damit stellt der Dice-Koeffizient ein Maß der Überlappung dar, was erklärt, warum er besonders bei Bildsegmentierungsproblemen eingesetzt wird und auch in der Prognose von Gehirngewebe bei Schlaganfällen als eine der wichtigsten Metriken etabliert ist. Ein Dice-Koeffizient von 0 bedeutet, dass kein Läsionsvoxel richtig vorhergesagt wurde. Der Wert 1 steht für das Ergebnis, dass die vorhergesagten Läsionsvoxel genau den tatsächlichen Läsionsvoxeln entsprechen. Im Vergleich zur ROC AUC, die im Anschluss an die Sensitivität und Spezifität beschrieben wird, fließen keine TNs in die Berechnung des Dice-Koeffizienten ein. Da Nichtläsionen beim Schlaganfall in größerer Anzahl vorkommen als Läsionsvoxel und daher einfacher vorherzusagen sind, können TNs bei einigen Metriken dazu führen, dass sie einen trügerisch positiven Eindruck des Modells vermitteln (siehe z. B. Accuracy).

Sensitivität und Spezifität

Die Sensitivität und Spezifität sind definiert als:

$$Sensitivität = \frac{|TP|}{|TP| + |FN|}$$

und

$$Spezifität = \frac{|TN|}{|TN| + |FP|}$$

Die Sensitivität entspricht damit dem Anteil an korrekt erkannten Läsionsvoxeln. Sie ist in Anbetracht der geringen Menge an Läsionsvoxeln besonders relevant. Die Spezifität benennt den Anteil an korrekt erkannten Nichtläsionsvoxeln und fällt bei der Schlaganfallprognose aufgrund der Vielzahl gesunder Voxel meist sehr hoch aus.

Passenderweise werden die Metriken auch als True-Positive- und True-Negative-Raten bezeichnet. Da bei einem Schwellwert von 0 alle Voxel als Läsionen (TP oder FP) klassifiziert werden, ist in diesem Fall die Sensitivität 1 und die Spezifität 0. Im Gegensatz hierzu werden bei einem Schwellwert von 1 (nahezu) alle Voxel als Nichtläsionen (TN oder FN) vorhergesagt, sodass in diesem Fall die Sensitivität 0 und die Spezifität 1 ist. Aufgrund dieser Gegenläufigkeit ist es wichtig, beide Metriken stets gemeinsam anzugeben.

Area under the receiver operating characteristic curve

Als weitere Metrik wird häufig die sogenannte area under the receiver operating characteristic curve, kurz ROC AUC betrachtet. Diese beschreibt die Fläche unter der ROC Kurve, welche den Verlauf der False-Positive-Rate (FPR) auf der x-Achse und den Verlauf der True-Positive-Rate (TPR) auf der y-Achse für alle Schwellwerte zwischen 0 und 1 beschreibt. Die FPR und die TPR sind definiert als

$$FPR = \frac{|FP|}{|FP|+|TN|} \text{ und } TPR = \frac{|TP|}{|TP|+|FN|}$$

Ein Verlauf entlang einer niedrigen FPR und einer hohen TPR deutet somit auf ein besonders vorteilhaftes Modell (siehe Abbildung 6). Diese Charakteristik spiegelt sich im Flächeninhalt der Kurve wider, die im besten Fall den Wert 1 annimmt.

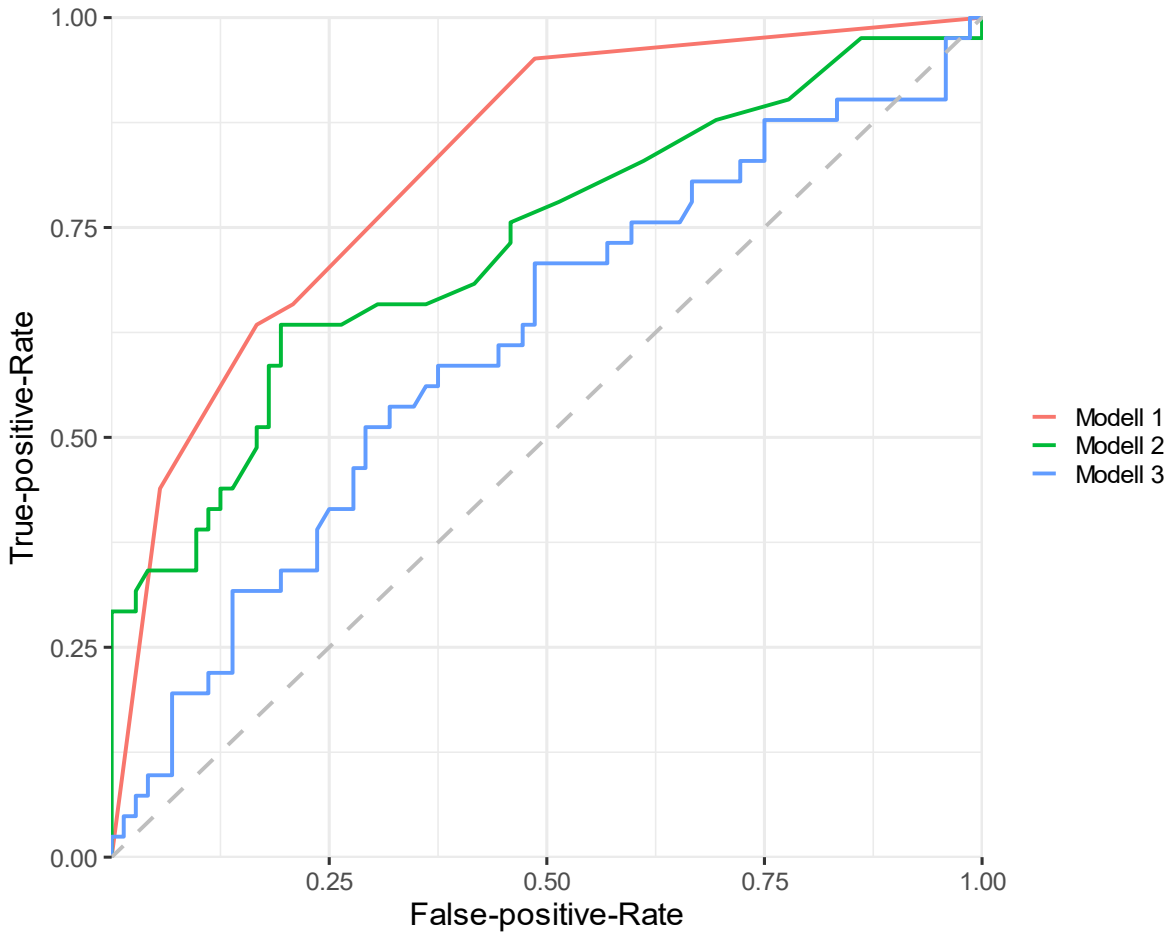

Abbildung 6: Beispielhafte ROC-Kurven für drei fiktive Modelle: Aus den Verläufen der ROC-Kurven der drei Modelle wird direkt ersichtlich, dass die Fläche unter der Kurve des ersten Modells größer ist als jeweils diejenige von Modell 2 und 3 (0,82 vs. 0,73 vs. 0,61). Ein Modell mit einer ROC AUC von 0,5 (gestrichelte Linie) wäre nicht besser als ein rein zufälliger Prognosescore. (Quelle: Dieses Beispiel wurde der Dokumentation des pROC-R-Pakets [Robin et al., 2011] entnommen und für diese Arbeit leicht modifiziert.)

Die ROC AUC lässt sich ebenfalls als die Wahrscheinlichkeit interpretieren, dass der Prognosescore für ein zufällig aus den Testdaten gezogenes Läsionsvoxel höher ist als für ein zufällig gezogenes Nichtläsionsvoxel. Somit beschreibt sie, wie gut ein Modell die unterschiedlichen Outcomes bei einer binären Klassifikation voneinander abgrenzt bzw. wie zuverlässig dieses die Voxel beim Schlaganfall nach ihrem Outcome sortiert. Mit einer ROC AUC von exakt 1 ist eine perfekte Vorhersage erreicht, bei der man durch einen Schwellwert alle Läsionsvoxel von den Nichtläsionsvoxeln unterscheiden könnte. Dagegen würde eine ROC AUC von 0,5 bedeuten, dass die Vorhersagen des Modells nicht besser sind als ein rein zufälliger Prognosescore. Da die ROC AUC von keinem Schwellwert abhängig ist, eignet sie sich besonders gut, um die Performance eines binären Klassifikationsmodells in einer Zahl zusammenzufassen. Daher wird sie häufig zum Vergleich verschiedener Modelle genutzt [Hosmer et al., 2013].

Als nächstes werden die in dieser Arbeit verwendeten Algorithmen zur Prognose des Gewebeschicksals erläutert, bevor die Grundlagen von neuronalen Netzen beschrieben werden, die eigenständig Features aus Bilddaten erlernen können und immer häufiger für die Läsionsvorhersage erprobt werden. Gleichzeitig wird aufgezeigt, inwiefern ein Problem bei der Erklärbarkeit von solchen Modellen im klinischen Setting besteht. Final wird der Forschungsstand zur modellbasierten Vorhersage von Gewebeoutcome beim akuten ischämischen Schlaganfall vorgestellt, aus dem sich die Forschungslücke hinsichtlich der räumlichen Faktoren, die in dieser Arbeit modelliert werden, ergibt.

2.5.2 Algorithmen

Logistische Regression

Die logistische Regression wurde in dieser Arbeit wegen ihrer Einfachheit, Recheneffizienz und ihrer Anwendung in mehreren früheren Studien verwendet wie u. a. [Schlaug et al., 1999; Wu et al., 2006; Jonsdottir et al., 2009; Kemmling et al., 2015; Flottmann et al., 2017]. Sie modelliert den Erwartungswert einer binären Zufallsvariable y in Abhängigkeit von verschiedenen unabhängigen Zufallsvariablen $x_1, x_2, \dots, x_n$ durch die Gleichung

$$E(y|x_1, x_2, \dots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Wenn die abhängige Variable als 1 (Läsion) und 0 (Nichtläsion) kodiert wird, dann entspricht der bedingte Erwartungswert der Wahrscheinlichkeit, dass ein Voxel eine Läsion entwickelt [Hosmer et al., 2013]. Diese Eigenschaft ist z. B. bei der Vorhersage des Infarktvolumens von besonderer Bedeutung, weil das Aufsummieren der Läsionswahrscheinlichkeiten über alle Voxel des Gehirns Prognosen darüber gestattet, welches Ausmaß das Infarktvolumen für einen Patienten annimmt [Flottmann et al., 2017]. Dies wurde z. B. von [Kemmling et al., 2015] genutzt, um Zusammenhänge zwischen der Rekanalisationszeit und dem Infarktvolumen zu erforschen.

Die Umkehrfunktion der logistischen Funktion entspricht dem Logarithmus des sogenannten Odds-Ratios, also dem Verhältnis der unterschiedlichen Wahrscheinlichkeiten für $y = 1$ vs. $y = 0$, und wird auch als Logit-Funktion bezeichnet. Für die resultierende Linearkombination der Features wird in dieser Arbeit der Begriff der Logit-Skala verwendet.

Die logistische Regression ist ein sehr effizienter Algorithmus. Ihre β-Koeffizienten ermöglichen es, den Einfluss von Änderungen in ihren Variablen (Features) auf den Prognosescore konkret zu beziffern. Die Haupteinschränkung der logistischen Regression besteht allerdings darin, dass nichtlineare Zusammenhänge und Interaktionseffekte explizit modelliert werden müssen, da sie diese nicht eigenständig aus den Daten erlernen kann [Hosmer et al., 2013]. Für das Training der logistischen Regression wird in dieser Arbeit die *glm*-Funktion aus dem in *Base-R* enthaltenen *stats*-Paket verwendet. Optimiert werden die Koeffizienten mit der Methode der iterativ neu gewichteten kleinsten Quadrate.

Random Forest

Die Wahl fiel auf den Random Forest-Algorithmus, da dieser Algorithmus gut mit kleinen Datenmengen wie den lokalen Trainingsdatensätzen umgehen kann und gleichzeitig effizient genug ist, um ihn auf einer größeren Zufallsauswahl (engl. sample) des gesamten Datensatzes anzuwenden. Zusätzlich erwies er sich in früheren Studien gegenüber der logistischen Regression als überlegen [Winder et al., 2019]. Ein Random Forest besteht aus einem Ensemble von Entscheidungsbäumen (engl. decision trees), die jeweils mit verschiedenen zufällig ausgewählten Daten trainiert werden. Einzelne Bäume sind typischerweise sehr anfällig für eine Überanpassung an die Trainingsdaten. Für die Anwendung auf neue Datensätze gilt eine Aggregation verschiedener unkorrelierter Bäume daher als robuster und besser generalisierbar.

Die finale Vorhersage eines Random Forests ergibt sich aus dem Durchschnitt aller Vorhersagen der einzelnen Entscheidungsbäume. Da die Bäume unabhängig voneinander trainiert werden, lässt sich das Training einfach parallelisieren. Dazu müssen allerdings sogenannte Hyperparameter (Einstellungen, die nicht aus den Daten erlernt werden können) wie u. a. die Anzahl der Bäume, ihre maximale Tiefe oder der Anteil an jeweils zufällig gezogenen Trainingsdaten spezifiziert werden.[19] Verringert sich dieser, dann führt dies zu einer geringeren Überlappung zwischen den Trainingsdaten der unterschiedlichen Bäume und damit zu schwächeren Korrelationen zwischen ihren Vorhersagen. Für die meisten Random-Forest-Implementierungen gilt allerdings, dass sie bereits für die Starteinstellungen relativ gute und robuste Ergebnisse erzielen [Probst et al., 2019].

Im Gegensatz zur logistischen Regression ist ein Random Forest in der Lage, nichtlineare Zusammenhänge und Interaktionen zwischen verschiedenen Variablen eigenständig zu lernen [Breiman, 2001]. Für das Training der Random-Forest-Modelle werden in dieser Arbeit die Implementierungen aus den *randomForest*- und *h2o*-R-Paketen genutzt (vgl. 3.5.2) [Liaw und Wiener, 2002].

XGBoost

Eine weitere auf Entscheidungsbäumen basierende Ensemble-Methode ist der Algorithmus XGBoost, der eine abgewandelte Implementierung der Gradient Boosting Machine darstellt. Eine Gradient Boosting Machine ist ein Ensemble, das aus mehreren schwachen Lernern (z. B. Entscheidungsbäumen) besteht, die nacheinander trainiert werden und so schrittweise die Prognosegüte der Gradient Boosting Machine verbessern. Hierzu werden vor jeder Trainingsrunde die Gewichte der einzelnen Trainingsdatenpunkte entsprechend den Residuen nach Abschluss der letzten Trainingsrunde angepasst.

XGBoost erweitert die Zielfunktion der Gradient Boosting Machine um einen Regularisierungsterm, welcher komplexere Entscheidungsbäume mit höheren Werten bestraft. Dies führt dazu, dass vor allem weniger komplexe Bäume in das Modell eingehen. Aus diesem Grund gilt XGBoost als besonders robust in Bezug auf Überanpas-

[19] Neben der Anzahl der Bäume, ihrer Tiefe und dem Anteil der zufällig gezogenen Trainingsdaten kann weiter spezifiziert werden, ob die Daten mit oder ohne Zurücklegen gezogen werden, wie viele Variablen zufällig für einen Split untersucht werden sollen, welche Split-Regel angewandt wird und wie viele Datensätze ein Knoten mindestens enthalten muss.

.

sungen an die Trainingsdaten. Mit verschiedenen Hyperparametern lässt sich der Lernprozess von XGBoost steuern, um damit das sogenannte Bias-Varianz-Dilemma weiter zu regulieren. In den vergangenen Jahren erregte dieser Algorithmus viel Aufsehen bei Machine-Learning-Wettbewerben, weil er sehr häufig zum Sieg führte [D. Nielsen, 2016]. Auch hinsichtlich der Trainingszeit bei tabellarischen Trainingsdatensätzen übertrifft XGBoost andere Machine-Learning-Modelle teilweise deutlich. Weitere Details zur Implementierung und der Optimierung von XGBoost wurden von [T. Chen und Guestrin, 2016] beschrieben. Für das Training der XGBoost-Modelle in dieser Arbeit werden die Funktionen aus dem gleichnamigen xgboost-R-Paket verwendet [T. Chen et al., 2017]. Auf die Einstellungen bzgl. der verschiedenen in dieser Arbeit erprobten Hyperparameterkombinationen wird näher in Abschnitt 3.5.2 eingegangen.

Künstliche neuronale Netze

Künstliche neuronale Netze (KNN) bieten eine ganze Reihe an Lernverfahren. Aufgrund ihrer Fähigkeit abstrakte Features aus unstrukturierten Daten zu erlernen, werden sie immer häufiger für Probleme aus der medizinischen Bildgebung erprobt. Im Folgenden werden einige ihrer grundlegenden Prinzipien vorgestellt. In Anlehnung an [Chollet und Allaire, 2018] werden der Aufbau eines vorwärtsgerichteten KNN und eines faltenden neuronalen Netzes (CNN, engl. convolutional neural network) beschrieben. Anschließend wird die Architektur eines U-Net vorgestellt, welches speziell für Segmentierungsprobleme entwickelt wurde und häufig in Studien zur Läsionsvorhersage verwendet wird [Ronneberger et al., 2015]. Eine theoretische Einführung in künstliche neuronale Netze bieten [Goodfellow et al., 2016], die die grundlegenden mathematischen Konzepte detailliert beschreiben.

Ein KNN besteht im Wesentlichen aus Knoten, welche innerhalb verschiedener Schichten angeordnet sind: einer Eingabeschicht für die Inputdaten, gefolgt von einer oder mehreren verdeckten Schichten (engl. hidden layers), in denen weitere Features erlernt werden, und einer Ausgabeschicht. In Anlehnung an die Signalübertragung im Gehirn werden die Knoten auch als künstliche Neuronen bezeichnet. Bei vollständig verbundenen Schichten sind alle Neuronen aus aufeinanderfolgenden Schichten über Kanten miteinander verbunden. Jede dieser Kanten wird durch ein initial zufällig gewähltes Gewicht beschrieben. Einzelne Neuronen werden über sogenannte Aktivie-

rungsfunktionen angeregt, was zumeist durch eine lineare oder eine logistische Funktion geschieht.

Als Argument einer Aktivierungsfunktion dient die gewichtete Summe aus der mit dem Neuron verbundenen Eingangswerten, welche um einen Biasterm ergänzt wird. Bei vorwärtsgerichteten Netzen fungieren die Neuronen aus der Eingabeschicht zusammen mit ihren Gewichten als Input für die Aktivierungsfunktionen der Neuronen aus der ersten verdeckten Schicht usw. Die Neuronen der letzten verdeckten Schicht dienen als Input für die Ausgabeschicht. Bei binären Klassifikationsproblemen kommt in dieser Schicht häufig die logistische Funktion zum Einsatz, sodass das Modell einen Wahrscheinlichkeitsscore zurückgibt.

Beim Lernen eines KNN werden die Ausgabewerte genutzt, um die Gewichte des KNN zu aktualisieren. Als Zielfunktion wird bei Klassifikationsproblemen überwiegend die Kreuzentropie verwendet, während die Gewichte üblicherweise mit dem sogenannten Backpropagation-Algorithmus aktualisiert werden. Dabei werden die Parameter in der Regel über das stochastische Gradientenabstiegsverfahren oder die adaptive Momentschätzung optimiert. Zur Regularisierung wird bei den Berechnungen zeitweise auf einen Teil der Neuronen verzichtet (Dropout). Dies limitiert die Relevanz einzelner Neuronen, sodass das Netz an Robustheit in der Anwendung in Bezug auf neue Datensätze gewinnt.

Die Anzahl der verdeckten Schichten wird auch als die Tiefe des KNN betrachtet. KNN mit einer hohen Anzahl an Schichten gelten als tiefe neuronale Netze (engl. deep neural networks) und werden zu den Methoden des tiefen Lernens (engl. deep learning) gezählt. Dem Bereich des oberflächlichen Lernens (engl. shallow learning) sind dagegen jene Methoden zuzuordnen, die keine abstrakten Merkmale aus ihren Eingaben lernen. Hierzu zählen z. B. die in dieser Arbeit verwendeten Algorithmen logistische Regression, Random Forest und XGBoost. Tiefe neuronale Netze zeichnen sich im Laufe des Trainings dadurch aus, dass in den verdeckten Schichten verschiedene Features vom Modell erlernt werden. Als äußerst vorteilhat erweist sich dies u. a. bei Machine-Learning-Problemen, bei denen unstrukturierte Daten wie z. B. Bild- oder Textdaten verwendet werden.

Im Bereich der Vorhersage des Gewebeschicksals beim Schlaganfall werden fast ausschließlich CNN eingesetzt, die aufeinander folgende Blöcke aus Faltungs- und Pooling-Schichten verwenden. Zur Aktivierung von Neuronen einer Faltungsschicht

werden die Bilddaten mithilfe eines Faltungskerns abgetastet und die Skalarprodukte zwischen Bildausschnitten und Faltungskernen berechnet, auf die noch ein Bias-Term addiert wird. Auf diese Weise wird die lokale Struktur der Bilddaten verarbeitet. Die Output-Werte einer Faltungsschicht stellen ein neues und ggf. verkleinertes Bild dar, dessen Elemente als Repräsentationen eines neu erlernten Features angesehen werden können. Üblicherweise kommen dabei mehrere Faltungskerne zum Einsatz, sodass mehrere Output-Bilder aus einer Faltungsschicht generiert (bzw. Features erlernt) werden. Die Auflösung der Features hängt von der Größe des Faltungskerns und dem Abtastschema ab. So können mit einem kleinen Faltungskern beispielsweise zuerst detaillierte Features wie Ecken und Kanten gelernt werden und anschließend Features, die auf Ecken und Kanten basieren. Um mit vergleichsweise kleinen Kerngrößen zunehmend größere Bildbereiche abzudecken und räumliche Hierarchien zu lernen, werden Pooling-Schichten eingebaut. In Pooling-Schichten werden die Output-Bilder der vorherigen Faltungsschicht mit einer konstanten Kerngröße abgetastet, wobei eine Aggregation wie z. B. das Maximum (Max-Pooling) oder der Durchschnitt (Average-Pooling) des Bildausschnitts gebildet wird. Im Anschluss an die Blöcke aus Faltungs- und Pooling-Schichten folgt eine Schicht, in der die Werte der Bilder in einen Vektor umgewandelt werden. Durch die Pooling-Schichten ist dieser Vektor hinreichend kurz, um eine oder mehrere vollständig verbundene Schichten einzubauen, auf die schließlich die Ausgabeschicht folgt.

Je nach Problemstellung können Elemente von neuronalen Netzen auf verschiedene Weisen miteinander kombiniert werden. So gibt es eine Reihe an bewährten Netzarchitekturen für verschiedene Aufgaben. Während CNN häufig zur Klassifikation von Bildern eingesetzt werden, verhilft dies in der biomedizinischen Bildverarbeitung vor allem dazu, ein Label pro Voxel zu annotieren, was eine räumliche Darstellung von Krankheitsbildern erlaubt. Dies wird als Segmentierung bezeichnet. Besonders geeignet ist die sogenannte U-Net-Architektur, weshalb diese auch häufig zur Läsionsvorhersage eingesetzt wird. Die einzelnen Schichten lassen sich bei dieser Architektur in zwei miteinander verbundenen gegenläufigen Pfaden anordnen, die sogenannten Encoder und Decoder (siehe Abbildung 7) [Ronneberger et al., 2015]. Die Rolle des Encoders ist es, Features aus den hochaufgelösten Bildinformationen zu bilden. Der Decoder bildet die Informationen aus den niedrig aufgelösten Feature-Karten dann wieder auf die Auflösung der Bildkarten ab [Badrinarayanan et al., 2017].

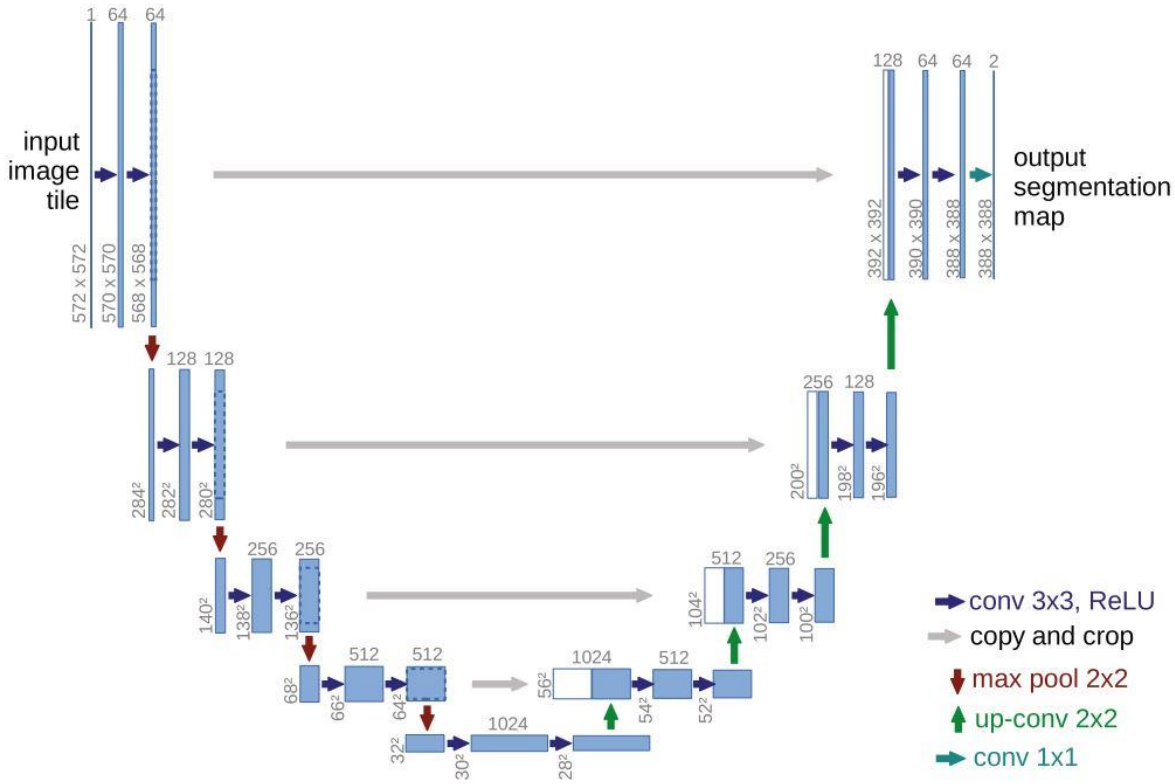

Abbildung 7: U-Net (Quelle: Ronneberger et al., 2015).

Beim U-Net beinhaltet der Encoder vier Blöcke (B1–B4) mit jeweils zwei Faltungsschichten. Die aufeinanderfolgenden Blöcke sind jeweils über eine Max-Pooling-Schicht verbunden, wodurch sich die Dimensionen der Output-Bilder von B1–B4 sukzessive verringern und Hierarchien zwischen den erlernten Features entstehen. Auf den Encoder folgt ein Block aus zwei weiteren Faltungsschichten (B5). Der Decoder besteht ebenfalls aus vier Blöcken (B6–B9), die jeweils eine Schicht zum Aneinanderhängen mehrerer Output-Werte sowie zwei darauffolgende Faltungsschichten umfassen. Jedem dieser Blöcke geht eine Upsampling-Schicht voraus, die zur Angleichung der Dimensionen der Input-Bilder für B6–B9 an die Dimensionen von B1–B4 führt. Zudem werden die Outputs der vier Encoder-Blöcke nicht verworfen, sondern im ersten Schritt in B6–B9 genutzt. So wird z. B. der Output aus B4 an den Input von B6 gehängt usw. Dadurch bedient sich die U-Net-Architektur sowohl lokaler Bildinformation als auch deren Kontext für die Segmentierung (bzw. Prädiktion). Bei der Vorhersage von ischämischen Läsionen ist der Einbezug solcher Informationen besonders hilfreich, da Läsionsgebiete oft zusammenhängen und ein Voxel in der Nähe des Infarktkerns auch selbst einem höheren Läsionsrisiko ausgesetzt ist.

Häufig wird bei der Entwicklung von Vorhersagemodellen die Frage, ob diese Modelle auch praxistauglich sind, vernachlässigt. Während die Einschätzung der Schlag-

anfallentwicklung auf Basis des DWI-PWI-Mismatches nachvollziehbaren Kriterien unterliegt, liefern komplexe binäre Klassifikationsmodelle keine Begründungen für ihre Gewebeprognosen mit. Sie entsprechen somit meist einer Blackbox. Ein Teilgebiet des Machine Learning beschäftigt sich daher mit der Erklärbarkeit von Modellen und deren Vorhersagen [Molnar, 2019].

2.5.3 Erklärbarkeit der Gewebeprognosen

Vor allem komplexe Verfahren, wie tree-basierte Ensemblemodelle und neuronale Netze, führen häufig zu präziseren Vorhersagen als gut interpretierbare Methoden, z. B. die logistische Regression [Gerke et al., 2020]. In der Medizin besteht schon aus ethischen und rechtlichen Aspekten eine besondere Notwendigkeit, automatisiert generierte Vorhersagen angemessen interpretieren zu können, weil die Interpretation erhebliche Konsequenzen für den Genesungsverlauf haben kann. Entscheidungsträger stehen deshalb vor einem Dilemma bei der Entscheidung darüber, welches Modell sie für ihre medizinischen Entscheidungen heranziehen: dasjenige mit einer potenziell besseren Vorhersage oder jenes mit einer guten Interpretierbarkeit [Wang et al., 2019; Zihni et al., 2020].

Ein Beispiel für diese Problematik zeigte sich erst kürzlich an einem von [Zihni et al., 2021] trainierten neuronalen Netz zur Vorhersage des funktionellen Outcomes nach einem Schlaganfall. Ein mRS von 0–2 entsprach dabei einem guten und ein mRS von 3–6 einem schlechten Outcome. Obwohl ihr Modell auf Basis des ADCs eine hohe ROC AUC von 0,92 auf den Testdaten aufwies, konnten sie zeigen, dass nicht etwa biologisch relevante Faktoren, sondern vor allem die Region am Rande des Gehirns hauptverantwortlich für die Prädiktionen waren. Dies erklärten die Autoren mit den Rahmenbedingungen bei der Bildgebung, z. B. den Bewegungen der Patienten oder Störungen des Magnetfeldes, die fälschlicherweise und zunächst unbemerkt vom Modell erlernt wurden.

Nach [Zihni et al., 2021] braucht es daher eine durch Fachwissen gestützte Interpretation von Modellen, die das Vertrauen von Ärzten, anderen medizinischen Fachkräften und Patienten in die Entscheidungsfindung bei Therapieverfahren stärkt. Hierzu muss vor allem die Interpretierbarkeit und sichere Überprüfbarkeit der Modelle während ihrer Entwicklung berücksichtigt werden.

In den aktuellen Leitlinien für die Behandlung des akuten ischämischen Schlaganfalls werden keine Empfehlungen zum Einsatz von Methoden der künstlichen Intelligenz gegeben. Da die Datenschutz-Grundverordnung 2018 europaweit in Kraft getreten ist, haben Patienten das Recht auf aussagekräftige Informationen über die Logiken, die hinter einer automatisierten Entscheidungsfindung auf Basis ihrer Daten stehen. Deshalb wird sich in Zukunft noch herausstellen, ob dazu eine Empfehlung in den Leitlinien stehen wird, die diese Anforderung erfüllt [Parliament, 2016].

2.5.4 Forschungsstand zur voxelbasierten Modellierung des Gewebeoutcomes

Überwiegend basieren binäre Klassifikationsmodelle zur Vorhersage des geschädigten Gewebes bei akuten ischämischen Schlaganfällen auf den initial erhobenen PWI- und DWI- Sequenzen, die auch für die schwellwertbasierte Mismatch-Bildung genutzt werden. Während die alleinige Kombination dieser Parameter mithilfe der logistischen Regression nicht automatisch zu signifikanten Verbesserungen gegenüber dem PWI-DWI-Mismatch führt [Wu et al., 2001], lassen sich mit zusätzlichen Features und geeignetem Algorithmus wesentlich bessere Ergebnisse erzielen [Winder et al., 2019; Benzakoun et al., 2021].

Ein Überblick über eine Vielzahl der bereits erprobten Methoden wird im Review von [Rekik et al., 2012] gegeben. Allein in sieben der insgesamt 14 vorgestellten Arbeiten wurde bereits damals der Rekanalisationsstatus berücksichtigt. Da Ischämien zusammenhängende Gebiete im Gehirn betreffen und sich mit der Zeit vom Kern aus auf einen Großteil des Perfusionsdefizits ausweiten, ist die Voxelnachbarschaft ebenfalls ein wichtiger Faktor für die Prognose des Gewebeoutcomes. Über Faltungsschichten lassen sich solche Nachbarschaftseffekte automatisch in Deep-Learning-Modellen erlernen, weshalb diese in den ISLES-Wettbewerben anstelle von Shallow Learning vermehrt eingesetzt wurden.

ISLES-Wettbewerbe

Nach einer Vielzahl weiterer Arbeiten zur Gewebeprognose (und zur Segmentierung) von Läsionsvoxeln wurde zur besseren Vergleichbarkeit von Methoden der automatisierten Segmentierung von Follow-up-Läsionen 2015 der ISLES-Wettbewerb (Ischemic Stroke Lesion Segmentation) ins Leben gerufen [Winzeck et al., 2018]. Die auto-

matisierte Segmentierung bietet ein wichtiges Hilfsmittel zur Berechnung des Läsionsvolumens, das häufig als Endpunkt in Schlaganfallstudien betrachtet wird. Darüber hinaus dienen die Segmentierungen der abgeschlossenen Infarkte auch als Ground Truth bzw. Target für das Training von Vorhersagemodellen. Deshalb verlagerte sich der Fokus der ISLES in den beiden Folgejahren (2016 und 2017) auf die Vorhersage von Schlaganfallläsionen.

Bei den Wettbewerben in den Jahren 2016 und 2017 traten 24 Teams gegeneinander an, um mit verschiedenen Strategien und Methoden das 90-Tage-Follow-up-Gewebeoutcome anhand von MRT-Daten für 19 (2016) bzw. 32 Patienten (2017) vorherzusagen. Dabei standen allen Teilnehmern dieselben Patientendatensätze als Trainingsdaten für die eingesetzten Algorithmen zur Verfügung (2016: n = 35 bzw. 2017: n = 43). Diese Datensätze beinhalteten ADC-, CBF-, CBV-, MTT- und T_{max}-Karten, die einheitlich und bewusst nur minimal vorprozessiert waren.

Während 2016 noch etwa die Hälfte der eingereichten Vorhersagen auf Methoden basierte, die keine abstrakten Features erlernen (Shallow Learning), entwickelten die Wettbewerbsteams 2017 ausschließlich Deep-Learning-Modelle. Die erhobenen Metriken umfassten die Präzision, die Sensitivität, die Hausdorff-Distanz, den durchschnittlichen symmetrischen Oberflächenabstand und den Dice-Koeffizienten. Die tatsächlichen Follow-up-Läsionen wurden für jeden Patienten von zwei unterschiedlichen Ratern manuell segmentiert und abschließend mit dem STAPLE-Algorithmus zusammengeführt (siehe [Warfield et al., 2004]).

Die Schwierigkeit einer genauen Läsionsprognose wird dabei schon aus der niedrigen Übereinstimmung der beiden Rater ersichtlich, für deren Segmentierungen der Follow-up-Läsionen aus der 2016er-Challenge ein durchschnittlicher Dice-Koeffizient von lediglich 0,58±0,20 erreicht wurde. Ebenfalls war das durchschnittliche Läsionsvolumen der Trainingsdaten für das Jahr 2017 mit 38 ml vergleichsweise klein gegenüber anderen Vorhersagestudien, was die Aufgabe erschwerte. Die Vorhersagemethode mit dem besten Gesamtergebnis im Jahr 2017 erreichte einen Dice-Koeffizienten von 0,31±0,23, und das beste Ergebnis allein in Bezug auf den Dice-Koeffizienten entsprach 0,32±0,23.

Um zukünftige Methoden mit den Ergebnissen aus den ISLES-Wettbewerben im Rahmen eines reproduzierbaren Workflows zu vergleichen, sind alle Datensätze aus den ISLES-Wettbewerben weiterhin online verfügbar. Diese konnten für die Zwecke

der vorliegenden Arbeit jedoch nicht verwendet werden, weil weder die Rohdaten der DWI- und PWI-Sequenzen verfügbar sind, noch die ADC- und Perfusionskarten vermutlich analog zu den Daten in dieser Arbeit prozessiert wurden, was potenziell zu einem systematischen Bias führen könnte. Deshalb werden in dieser Arbeit einzig Daten aus dem wesentlich größeren I-KNOW-Datensatz genutzt (Abschnitt 3.1).

Obwohl sich die Vorhersagequalität der entwickelten Methoden kontinuierlich in jedem Wettbewerbsjahr verbesserte, blieb ein grundlegender Fortschritt noch aus. Selbst mit einer gestiegenen Menge an verfügbaren Trainingsdaten im Jahr 2017 scheinen die Vorhersageergebnisse durch das komplexe Geschehen eines Infarkts bisher limitiert.

Aus den ISLES-Wettbewerben lässt sich dennoch die Erkenntnis ziehen, dass aktuelle und neu entwickelte Methoden neben klinischen Informationen auch a priori bekannte physiologische Informationen zu den Gehirnläsionen einbeziehen müssen und gleichzeitig die Transparenz und Interpretierbarkeit der Modelle gewährleistet sein muss. Bei allen Modellen fehlt bislang die Berücksichtigung der Voxellage, obwohl die Versorgungsgebiete der einzelnen Gehirnarterien räumlich beschränkt sind und die räumliche Verteilung der Läsionen bei einem Gefäßverschluss nicht zufällig ist. So konnten [Shen und Duong, 2008] bei einer Studie an Ratten nachweisen, dass die räumliche Lageinformation entscheidend für die Läsionsvorhersage ist. [Kemmling et al., 2015] belegte den Einfluss der räumlichen Lage auf die Läsionsvorhersage für den Menschen.

Diese Erkenntnisse werden im Forschungsdesign dieser Arbeit insofern berücksichtigt, als räumliche Features in die Modellierung implementiert und Metriken zur Validierung der Vorhersageergebnisse (Shapley-Werte) vorgeschlagen werden.

Deep-Learning zur Integration von Voxelnachbarschaften

Gemäß Neuausrichtung der ISLES-Wettbewerbe 2016/17 wurde vermehrt an tiefen CNN-Modellen im Zusammenhang mit Schlaganfällen geforscht. Ihre größte Stärke hinsichtlich dieser Problemstellung ist, dass sie Voxel nicht unabhängig voneinander betrachten, sondern über ihre Faltungsschichten automatisch komplexe räumliche Wechselwirkungen innerhalb von Voxelnachbarschaften hierarchisch modellieren. Aufgrund der Heterogenität von Schlaganfällen ist davon auszugehen, dass sich die Vorhersagen von CNN in besonderem Maße mit einer steigenden Anzahl an Trainings-

datensätzen verbessern werden. Zur Verfeinerung der bisherigen Ansätze wurden u. a. Verfahren zum Einbezug der aufeinanderfolgenden PWI-Aufnahmen [Pinto et al., 2018b] entwickelt, die komplementär zu den PWI-Parametern die Zusammenhänge zwischen akuten Blutflussdynamiken und dem Gewebeoutcome abbilden. Zur standardisierten Auswahl der PWI-Karten verwendeten sie ein symmetrisches Zeitfenster, um den Zeitpunkt des maximalen Kontrastmittels. Diese vergleichsweise einfache Integration der PWI-Karten führte zu einem starken weiteren Anstieg des Dice-Koeffizienten von 0,20 auf 0,29, welches ein Gesamtergebnis ist, das im oberen Bereich der ISLES-Einreichungen liegt.

Weiter wurde der Einfluss einer Rekanalisation integriert, denn ihre (frühe) Herbeiführung bietet nachweislich den größten therapeutischen Hebel zur Minimierung des Gewebeschadens. Dementsprechend verfolgten [Pinto et al., 2018a] anhand der frei verfügbaren ISLES-Datensätze den Ansatz, die Zielfunktion für das Training eines CNN zu modifizieren, indem falsch vorhergesagte Läsionsvoxel bei Patienten mit Rekanalisation (hoher TICI-Score) weniger stark gewichtet wurden als bei Patienten ohne oder nur minimaler Rekanalisation. Ihr finales Modell, das weiterhin klinische Metadaten beinhaltet, verbesserte sich durch den Rekanalisationsstatus im Dice-Koeffizienten um 0,05 auf 0,29 und liegt damit ebenfalls im Bereich der Top-Ergebnisse der ISLES.

Allerdings wurden auch hier keine räumlichen Informationen genutzt, obwohl [Shen und Duong, 2008] bereits im Tierversuch an Ratten gezeigt haben, dass der Einfluss der Lage insbesondere bei einer gelungen Rekanalisation an Bedeutung gewinnt. Diesen Zusammenhang führten [Shen und Duong, 2008] auf eine nicht zufällige Verteilung der kollateralen Gefäßstrukturen zurück, die für einen möglichst langen Erhalt der Penumbra ausschlaggebend sind, sodass die Bedeutung einer erfolgreichen Rekanalisation für das Gewebeoutcome räumlich variiert. Gerade beim Menschen mit seiner weitaus komplexeren Gehirnstruktur, erwarten die Autoren der Tierstudie daher noch weitaus größere positive Effekte durch den Einbezug der Voxellage als bei Ratten.

Zwei Ergebnisse auf größeren Datensätzen wurden rein auf der Basis von Biomarkern erzielt: So veröffentlichten [A. Nielsen et al., 2018] ein CNN zur Vorhersage des 30-Tage-Follow-up-Gewebeoutcomes auf einer modifizierten Variante der sogenannten SegNet-Architektur [Badrinarayanan et al., 2017]. Mit 158 MRT-Patientendatensätzen trainierten sie ein 37 Schichten umfassendes Netz anhand von neun Biomarkern. Das Modell erreichte bei der Validierung an einem Testdatensatz aus 29 Patienten eine ROC AUC von 0,88±0,12. Von ursprünglich 222 verfügbaren Datensätzen

entstammten 105 dem auch in dieser Arbeit verwendeten I-KNOW-Datensatz (vgl. Abschnitt 3.1). Während ihr Modell signifikante Verbesserungen von über 0,10 in der ROC AUC gegenüber einer logistischen Regression sowie schwellwertbasierten Einteilungen (ADC und T_{max}) aufwies, zeigte sich diesbezüglich ein eher geringer Effekt von 0,03 im Vergleich zu einem Shallow-Network, das keine Faltungsschichten und damit keine Nachbarschaftsinformationen beinhaltet. Ein Vorteil der Faltungsschichten des CNN lag darin, dass die Vorhersagekarten weniger rauschanfällig waren als die einer logistischen Regression und einer schwellwertbasierten ADC-Vorhersage. Ein Nachteil der Faltungsschichten ist jedoch die sehr lange Trainingszeit von fünf Tagen für das CNN.

[Yu et al., 2020] trainierten ein CNN (U-Net-Architektur) an 182 Patientendatensätzen aus zwei Studien: der iCAS- und der DEFUSE-2-Studie. Dabei erreichte das auf den üblichen DWI- und PWI-MRT-Parametern trainierte Modell bei der Prädiktion der 3 bis 7-Tage-Follow-up-Läsionen im Rahmen einer Kreuzvalidierung eine durchschnittliche ROC AUC von 0,89 und einen durchschnittlichen Dice-Koeffizienten von 0,53. Der vergleichsweise sehr hohe Dice-Koeffizient wurde von den Autoren vor allem auf die relativ hohe Anzahl an verfügbaren Trainingsdatensätzen zurückgeführt. Anzumerken ist allerdings, dass der Median Dice-Koeffizient in der Subgruppe der Patienten mit minimaler Reperfusion (bei durchschnittlichem FU-Läsionsvolumen von 86 ml) mit 0,58 deutlich höher gegenüber dem Wert von 0,48 für Patienten mit vollständiger Reperfusion (durchschnittliches FU-Läsionsvolumen von 23 ml) ausgefallen ist. Der Median des FU-Läsionsvolumens betrug 54 ml. In verschiedenen Studien erreichte der Dice-Koeffizient für große Läsionen bessere Werte, da diese einfacher als kleine Läsionen vorherzusagen sind [Winder et al., 2019; Benzakoun et al., 2021].

Nachbarschaftsmodellierung bei Shallow Learning

Zwar arbeiteten Wissenschaftler bei Shallow-Learning-Methoden weniger an der Integration von Nachbarschaftsinformationen, aber auch dort gab es vereinzelt (erfolgreiche) Versuche, die direkte Nachbarschaft in die Modellierung zu integrieren, um die Vorhersagen der Läsionswahrscheinlichkeit zu verbessern.

So verfolgten beispielsweise [Nguyen et al., 2008] den Ansatz, die räumliche Korrelation zwischen dem Infarktoutcome benachbarter Voxel über ein Spatial Autoregressive Model zu berücksichtigen, wozu sie einen rekursiven Korrelationsterm in eine logistische Regression einbauten, der die sukzessive Ausbreitung eines Infarkts über

benachbarte Voxel modelliert. Im Rahmen einer LOPO-Kreuzvalidierung an 79 Patienten erreichte ihr Modell mithilfe des ADC, der T2 und verschiedenen PWI-Karten eine signifikante Steigerung der ROC AUC von 75 auf 79 %. Eine Limitation ihres Modells ist allerdings, dass sie aufgrund großer Abstände zwischen den Schichten nur Nachbarschaften innerhalb derselben Schicht modellierten.

Eine allgemeinere Methode zur Nachbarschaftsmodellierung auf Basis von rezeptiven Feldern wurde zum ersten Mal von [Scalzo et al., 2012] zur Schlaganfallprognose angewandt. Ihre Methode basiert im Wesentlichen darauf, Features innerhalb kuboidförmiger Voxelnachbarschaften in die Prädiktion einzubeziehen, weswegen dieser Ansatz hier als patchbasierte Nachbarschaftsmodellierung bezeichnet wird. Als Features verwendeten sie die ADC- und T_{max}-Karten von 25 Schlaganfallpatienten. Aufgrund der Vervielfältigung der Features reduzierten sie die Trainingsdaten auf 10.000 Voxel. Da benachbarte Datensätze stark miteinander korrelieren, überrascht es kaum, dass sie mit einem nichtlinearen Modell (Kernel Spectral Regression) bessere Ergebnisse gegenüber einer logistischen Regression erzielten, wobei die Ergebnisse des nichtlinearen Modells zusätzlich mit zunehmender Nachbarschaftsgröße stabil blieben. So konnten die ROC AUC für das T_{max}-basierte Modell von 83,5 auf 90,9 % und für das ADC-basierte Modell von 81,9 auf 86,7 % gesteigert werden. Auch in der Studie von [Scalzo et al., 2012] war es eine Limitation, dass Voxelnachbarschaften dort lediglich schichtweise modelliert wurden.

Ähnliche Ergebnisse erhielten allerdings auch [Klug et al., 2020], die diesen Ansatz anhand von 144 CT-PWI-Datensätzen für dreidimensionale Patches ebenfalls mithilfe der logistischen Regression erprobten. Sie erzielten damit sowohl für ein T_{max}-Modell als auch für ein Modell mit mehreren PWI-Parametern eine Steigerung der ROC AUC von 0,79 auf 0,89 sowie des Dice-Koeffizienten von 0,11 auf 0,16. Analog zu den Vorhersagkarten der CNN wiesen auch die Ergebnisse der Shallow Learner bei der Integration von Nachbarschaftsvoxeln ein reduziertes Rauschen auf.

Eine weitere Methodik zur Integration patchbasierter Features nutzten [McKinley et al., 2016]. So berechneten sie unterschiedliche Statistiken, wie den Mittelwert, den Median und verschiedene Perzentile der Bildgebungsparameter (auch als lokale Histogramme bezeichnet) innerhalb unterschiedlich großer Patches. Zusammen mit dem Rekanalisationsstatus und der zugehörigen Rekanalisationszeit führten die Features zu einem vierten Platz ihres Random Forests in der ISLES-Challenge von 2016.

Aktuelle tree-basierte Ansätze zur Gewebeprognose

Abgesehen von Nachbarschaftsmodellierungen wurde nach den ISLES-Wettbewerben zum ersten Mal von [Livne et al., 2018] der XGBoost-Algorithmus zur Vorhersage des Schlaganfalloutcomes getestet, und [Winder et al., 2019] griffen bei der Suche nach optimalen Voraussetzungen zum Training verschiedener Algorithmen auf Features wie den Abstand eines Voxels zum Infarktkern und den Gewebetyp (Wahrscheinlichkeit für graue Substanz) sowie die Gehirnregion zurück.

Nach der Veröffentlichung des XGBoost-Algorithmus im Jahr 2014 hat dieser sich bereits in mehreren Wettbewerben bewährt, dennoch dauerte es noch vier Jahre bis zu seiner ersten Anwendung für die Schlaganfallprognose: [Livne et al., 2018] publizierten dazu in „Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke“. Informationen zur Lage oder Nachbarschaft der Voxel wurden jedoch nicht verwendet, und auch die Zeitpunkte der Follow-up-Aufnahmen wurden nicht angegeben. Trotz einer hohen ROC AUC von 0,88 wird XGBoost für die Schlaganfallprognose immer noch selten angewendet. Dies ist ein Grund, weswegen in der vorliegenden Arbeit für die Modellierung XGBoost als Algorithmus ausgewählt wurde (vgl. Abschnitt 3.5.2).

Eine ausführliche Validierung von verschiedenen Shallow-Learning-Algorithmen, z. B. dem ebenfalls tree-basierten Random Forest, der logistischen Regression und dem K-Nearest-Neighbor-Algorithmus, haben [Winder et al., 2019] durchgeführt. Als Metrik wurde dabei der Dice-Koeffizient betrachtet. Zur Integration der Voxelnachbarschaft wurde für jedes Voxel die Distanz zu einem vorher definierten akutem Infarktkern ($ADC < 550 \cdot 10^{-6}$ mm^2/s) berechnet.

Alle Random-Forest-Modelle zeigten sich gegenüber den anderen Algorithmen und der besten schwellwertbasierten Einteilung des Läsionsgewebes anhand des ADC oder eines einzelnen PWI-Parameters überlegen. Demzufolge sollten in zukünftigen Studien Random Forests als Vergleichsmethodiken für Deep-Learning-Modelle herangezogen werden. Für das Training erwies sich ein ausbalanciertes Downsampling der beiden Voxelklassen pro Patienten als vorteilhaft für alle Modelle. Demgegenüber brachte eine Normalisierung der PWI- und DWI-Parameter keine Vorteile für die Prognosen.

Für eine separate Modellierung des Rekanalisationsstatus wurden die Patienten in drei Kohorten aufgeteilt: nicht-rekanalisiert (n = 39) und erfolgreich rekanalisiert durch

Thrombolyse (n = 23) bzw. Thrombektomie (n = 38). Erwartungsgemäß fiel der Dice-Koeffizient für Patienten mit Rekanalisation am niedrigsten aus, denn zum einen entwickelten sich bei ihnen die kleinsten Infarkte und zum anderen erschwert eine erfolgreiche Rekanalisation die Problemstellung, da ab dem Zeitpunkt der Rekanalisation die Werte aus den initialen DWI- und PWI-Aufnahmen ihre Gültigkeit verlieren. Der positive Effekt einer Rekanalisation hängt jedoch gerade bei einer Rekanalisation vom Ort des Verschlusses bzw. der kollateralen Blutversorgung für das Versorgungsgebiet der verstopften Arterie ab. Nach [Shen und Duong, 2008] dient folglich der Einbezug einer detaillierten räumlichen Läsionsverteilung, basierend auf Verschlüssen für dasselbe Stromgebiet, als Indikator für eine kollaterale Durchblutung.

Anstelle eines redundanten Features, das für jedes Voxel eines Patienten dessen Rekanalisationsstatus beinhaltet, wurde ein Modell pro Kohorte trainiert. Dies wurde damit begründet, dass ansonsten durch eine Überlappung dieser unterschiedlichen Verlaufsindikatoren in den Trainingsdaten, welche möglicherweise nicht richtig vom Modell erlernt werden, ein systematisches Bias auftreten könnte. Allgemein ist die Integration von derartigen Features auf Patientenebene problematisch, da sich die Information von einem Patienten auf etwa 100.000 Trainingsvoxel[20] repliziert und somit einen unverhältnismäßig großen Anteil an der Trainingsdatenmenge ausmacht. Dieses Problem ist auch bei der Integration weiterer Patientenfeatures Alter und Geschlecht, dem Zeitabstand zwischen ersten Symptomen und Bildgebung oder klinischen Scores, wie dem mRS und NIHSS, relevant. Und so tritt es auch in anderen Arbeiten wie z. B. [Kemmling et al., 2015] oder [Pinto et al., 2018a] auf.

Integration von räumlichen Informationen

Während Konzepte für die Integration von Nachbarschaftsinformationen für alle Modelltypen existieren, ist ein typischer Nachteil der bisher vorgestellten Arbeiten, dass bisher kaum räumliche Informationen über die Voxel in die Modellierung des Gewebeoutcomes aufgenommen wurden, obwohl die räumliche Verteilung von Läsionen nicht zufällig auftritt und bereits 2008 vielversprechende Ergebnisse durch den Einbezug von räumlichen Informationen bei Ratten erzielt wurden.

[20] Die genaue Zahl variiert u. a. mit der Auflösung der verwendeten Karten und dem Sampling der Trainingsdaten. Bei der Verwendung eines MNI-Atlas kann sie sich ggf. stark erhöhen.

Als Erste integrierten [Shen und Duong, 2008] räumliche Informationen in die Vorhersage des Gewebeschicksals beim ischämischen Schlaganfall im Tierversuch. Für drei Gruppen von jeweils zwölf Ratten mit unterschiedlichem Rekanalisationsstatus (30 Minuten, 60 Minuten, permanent) ermittelten sie Infarktwahrscheinlichkeiten auf Voxelebene, deren Wert auf dem ADC und dem CBF von jeweils der Hälfte der Tiere basierte. Die Testvorhersagen für die jeweils anderen sechs Tiere zeigten jedoch noch keine ausreichende Übereinstimmung mit dem tatsächlichen Gewebeoutcome. Aufgrund einer heterogenen Infarktverteilung für Ratten[21] wurde ein gewichtetes Mittel aus Infarktvorhersagen und räumlicher Infarktwahrscheinlichkeit gebildet, die als relative Infarkthäufigkeit pro Voxel berechnet wurde. Die so erhaltene Prädiktion wies signifikante Verbesserungen auf und führte innerhalb der verschiedenen Rekanalisationsgruppen zu ROC AUC von 0,87 (30 Minuten), 0,90 (60 Minuten) und 0,93 (permanente Verschlüsse).

Der optimale Anteil der räumlichen Vorhersage am gewichteten Mittel variierte je nach Rekanalisationsstatus zwischen 0,1 (keine Rekanalisierung) und 0,4 (Rekanalisierung nach 30 bzw. 60 Minuten). Zusätzlich variierten die relativen Infarktkarten erheblich zwischen den einzelnen Rekanalisationsgruppen. Die Autoren der Studie vermuteten einen deutlich größeren Anteil der räumlichen Effekte bei Vorhersagemodellen für Menschen, da die Ausprägungen eines Schlaganfalls im menschlichen Gehirn noch heterogener sind als bei Ratten. Darüber hinaus sollte für diese auch der Gewebetyp (weiß vs. grau) in die Prädiktion einbezogen werden, weil sich daraus unterschiedliche Normwerte für die Perfusion ergeben.

Erst [Kemmling et al., 2015] griffen diese Aspekte beim Menschen auf und integrierten eine Wahrscheinlichkeitskarte für Läsionen bei Verschlüssen im vorderen Bereich der MCA in ihrer Arbeit, bei der sie den Einfluss der Zeit auf das Gewebeoutcome bis zur Rekanalisation basierend auf einem Vorhersagemodell untersuchten. Hierzu wählten sie eine logistische Regression mit den folgenden Einflussfaktoren:

1. PWI-Features: CBF, CBV, MTT, TTD (engl. time to drain; Zeit bis zum Washout des Kontrastmittels), räumliche Cluster von niedrigen CBV- und TTD-Werten und binäre Karten für Positionen, in denen keine endlichen Werte für MTT und TTD bestimmt werden konnten,

[21] Z. B. ist ihr Gewebe in der Nähe der vorderen und hinteren Gehirnschlagader aufgrund der anhaltenden oder partiellen kollateralen Durchblutung weniger anfällig für ischämische Schädigungen (das Gegenteil gilt für den Hippocampus).

2. Rekanalisations-Features: Zeit bis zur Bildgebung und Rekanalisation, Rekanalisationsstatus, Interaktion zwischen der Zeit bis zur Rekanalisation und dem Rekanalisationsstatus,
3. dem Gewebetyp,
4. Läsionsverteilung für einen externen Datensatz von 112 Patienten mit Gefäßverschlüssen im vorderen MCA-Bereich,
5. weitere klinische Variablen wie das Alter, das Geschlecht und den NIHSS.

Ihr Modell, das auf CT-Datensätzen von 161 Schlaganfallpatienten trainiert wurde, erzielte im Rahmen einer LOPO-Kreuzvalidierung eine ROC AUC von 0,85. Abgesehen von einem negativen Modellkoeffizienten für die Wahrscheinlichkeitskarte wurde festgestellt, dass durch die MCA-Läsionskarte Voxel mit Perfusionsdefiziten außerhalb der Versorgungsgebiets der MCA als gesundes Gewebe vorhergesagt wurden.

Da die absolute Wahrscheinlichkeit allerdings kein alleiniger Faktor für die Entwicklung einer Läsion im MCA-Gebiet selbst ist, und die Ergebnisse von [Shen und Duong, 2008] nahelegen, dass die Wahrscheinlichkeit zusammen mit den akuten Durchblutungsparametern betrachtet werden muss, fehlen für eine adäquate Modellierung diesbezüglich Interaktionsterme. Diese müssten bei der logistischen Regression explizit konfiguriert werden, wohingegen (nichtlineare) Interaktionseffekte in tree-basierten Modellen, die sich in bisherigen Studien als präziser für die Vorhersage des Gewebeschicksals herausstellten [Winder et al., 2019], automatisch vom Modell erlernt werden.

Erst [Benzakoun et al., 2021] verbinden die Nutzung von regionalen Nachbarschaftsinformationen und der Lage der Voxel und nutzen überdies die Werte der kontralateralen Nachbarschaften als Features in ihren tree-basierten Modellen. Sie trainierten je zwei XGBoost-, Random-Forest- und CNN- (U-Net-Architektur) Modelle zur Vorhersage des 24 h FU-Infarkts auf einem Datensatz von 394 Patienten. Einmal wurden die Modelle anhand der Parameter ADC und T_{max} und einmal anhand von ADC, CBF, CBV, MTT und T_{max} trainiert. Zur Nachbarschaftsmodellierung zogen die Wissenschaftler für die tree-basierten Modelle analog zu [Scalzo et al., 2012] und [Klug et al., 2020] einen auf kuboid-förmigen Patches basierenden Ansatz heran, wobei für jedes Voxel die Parameter der umliegenden Voxel (innerhalb eines $5 \times 5 \times 3$-Patches) als zusätzliche Features verwendet wurden. Ebenso sind die ADC-Werte des korrespondierenden kontralateralen $5 \times 5 \times 3$-Patches und die MNI-Atlas-Positionen der einzelnen Voxel enthalten. Trotz der hohen Anzahl an Features war das Training der

XGBoost- (89 Minuten) und Random-Forest- (19 Minuten) Modelle deutlich schneller beendet als das Training des CNNs (160 Minuten). Sowohl bei der Verwendung von lediglich T_{max} als auch beim Einsatz aller PWI-Parameter waren die Mediane der Dice-Koeffizienten von XGBoost (0,53 bzw. 0,54) und Random Forest (0,52 bzw. 0,51) höher als die Werte der CNN-Modelle (je 0,48). Dabei waren die Unterschiede zwischen den XGBoost- und den CNN-Modellen signifikant. Für die CNN-Modelle zeigten sich bzgl. des Dice-Koeffizienten keine signifikanten Vorteile gegenüber einem als T_{max} (> 6s) und ADC ($< 620 \cdot 10^{-6}$ mm^2/s) definierten PWI-DWI-Mismatch (0,45).

In der Gruppe mit bekanntem Rekanalisationsstatus (n = 87) führte dessen Berücksichtigung durchgehend zu signifikanten Verbesserungen des Dice-Koeffizienten: XGBoost (0,55 vs. 0,47), Random Forest (0,48 vs. 0,46), U-Net (0,46 vs. 0,38), PWI-DWI-Mismatch[22] (0,49 vs. 0,39).

Der insgesamt sehr hohe Dice-Koeffizient wurde in ihrer Arbeit, neben den methodischen Fortschritten, auch mit der hohen Patientenzahl und insbesondere mit den großen FU-Läsions-Volumina von durchschnittlich 60 ml begründet. Es ist bekannt, dass der Dice-Koeffizient häufig höhere Werte für große Läsionen aufweist. Dieser Effekt wurde verstärkt für CNN-Modelle beobachtet, weshalb diesbezüglich angeführt wurde, dass die Schichtabdeckung des CNN (32×32) weit größer war als bei den tree-basierten Modellen (5×5), die sich deshalb eher für die Vorhersage von großen Läsionen eignen müsste. Zudem sind die von CNN erlernten Features translationsinvariant, d. h. sie können mit gleicher Wahrscheinlichkeit an unterschiedlichen Positionen auftreten. Diese Modellierung ist für den Schlaganfall nicht angemessen, da die Läsionen durch die Versorgungsgebiete der verschlossenen Gefäße eingeschränkt sind, sodass sie nicht zufällig verteilt sind. Hier werden die MNI-Positionen als Features in den tree-basierten Modellen relevant, anhand derer die Ensembles z. B. erlernen, ob die Nachbarschaft innerhalb oder außerhalb desselben Territoriums liegt oder sich über mehrere Versorgungsgebiete erstreckt.

Eine Auswahl der hier vorgestellten Arbeiten zur Vorhersage des Gewebeoutcomes beim Schlaganfall ist in Tabelle 1 zusammengefasst.

[22] In diesem Zusammenhang ist das PWI-DWI-Mismatch bei Rekanalisierung als ADC-Läsion und bei Nichtrekanalisierung als ADC-Läsion + Perfusionsdefizit definiert.

Tabelle 1: Ausgewählte Veröffentlichungen zur Vorhersage des Gewebeoutcomes beim akuten ischämischen Schlaganfall

Autoren/ Jahr	Methode[A]	Features	Anzahl Patienten (Training/Test)	Ø FU-Läsionsvolumen (ml)	Ø ROC AUC	Ø Dice-Koeffizient
[Kemmling et al., 2015]	Logistische Regression	PWI- u. Rekanalisations-Features, Gewebetyp, Läsionswahrscheinlichkeiten, klinische Informationen	161[B]	68,2	0,85	-
[Pinto et al., 2018a]	Deep Learning (U-Net)	DWI- u. PWI-Features, Rekanalisationsstatus u. klinische Metadaten	75 (43/32)	36	-	0,29
[Pinto et al., 2018b]	Deep Learning (U-Net + GRU)	DWI- und PWI-Features inkl. roher PWI-Karten	75 (43/32)	36	-	0,35
[A. Nielsen et al., 2018]	Deep Learning (SegNet)	Neun MRT-Biomarker	187 (158/29)	-	0,88	-
[Livne et al., 2018]	XGBoost	ADC, FLAIR und 10 PWI-Features	195[B]	-	0,88	-
[Winder et al., 2019]	Random Forest	ADC u. PWI-Features, Gehirnregionen, Gewebetyp, Abstand zum Infarktkern, klinische Informationen	39[B]/23[B]/38[B] (je nach Rekanalisationsstatus)	-	-	0,45
[Yu et al., 2020]	Deep Learning (U-Net)	DWI- und PWI-Features	182[C]	54[E]	0,89	0,53
[Klug et al., 2020]	Logistische Regression	PWI-Features in $7 \times 7 \times 7$-Patches	144[C]	-	0,89	0,16
[Benzakoun et al., 2021]	XGBoost	DWI- und PWI-Features in $5 \times 5 \times 3$-Patches (ipsi- und kontralateral) sowie MNI-Koordinaten	394[D]	60	0,94[E]	0,53[E]
	Deep Learning (U-Net)	DWI- und PWI-Features			0,94[E]	0,48[E]

[A] Beste Methode pro Veröffentlichung

[B] Leave-One-Patient-Out-Kreuzvalidierung

[C] 5-Fold-Kreuzvalidierung

[D] 10-Fold-Kreuzvalidierung

[E] Median

2.5.5 Forschungslücke

Die Auswertung der bisherigen Arbeiten ergibt, dass die Modellierung auf Basis der Bildgebungsparameter bereits weit fortgeschritten ist. Dabei werden Nachbarschaftsinformationen, z. B. bei neuronalen Netzen, über die Faltungsschichten in CNN integriert und können alternativ patchbasiert als Features benachbarter Voxel oder über den Abstand eines Voxels zum Infarktkern modelliert werden.

Räumliche Faktoren wurden dagegen, trotz guter Ergebnisse beim Rattenschlaganfall [Shen und Duong, 2008], kaum für die Vorhersage beim Menschen aufgegriffen und tauchten eher als Nebenprodukte in anderen Studien auf ([Kemmling et al., 2015]). Weder wurden die Effekte gesondert validiert noch wurden Empfehlungen für den gezielten Einsatz von räumlichen Features ausgesprochen. So wurden relative Läsionshäufigkeiten zur Outcomevorhersage beim Menschen bislang lediglich im Rahmen der logistischen Regression anhand von CT-Daten bei [Kemmling et al., 2015] mit einer Vielzahl weiterer Features zur Modellierung des Effekts der Rekanalisationszeit auf das Infarktvolumen genutzt. Interaktionen zwischen den Läsionshäufigkeiten und den anderen Features wurden nicht explizit modelliert. Dadurch kann es passieren, dass Gebiete mit hoher/niedriger Läsionswahrscheinlichkeit die individuellen Indikatoren aus der Bildgebung überdecken und es zu starken Überanpassungen der Modelle kommt.

Neben der Eigenschaft von tree-basierten Modellen Interaktionsterme selbstständig zu erlernen, ist anzunehmen, dass räumliche Features aufgrund der komplexen anatomischen Strukturen des menschlichen Gehirns wesentlich besser in diesen nichtlinearen Modellen zur Geltung kommen. Allerdings wurde auch für die MNI-Koordinaten, die von [Benzakoun et al., 2021] zusammen mit einer patchbasierten Nachbarschaftsmodellierung in einem Random-Forest- und in einem XGBoost-Modell verwendet wurden, bislang keine eigenständige Validierung hinsichtlich eines Effektes dieses räumlichen Features durchgeführt, sodass für den Einsatz im Klinikalltag keine Empfehlung gegeben werden konnte.

Abgesehen von der Integration räumlicher Informationen in globale Modelle, mangelt es an einem angemessenen Modellierungskonzept, welches die Integration von Patientenfeatures, wie des Rekanalisationsstatus und NIHSS etc., in globale Modelle erlaubt. Dies hat dazu geführt, dass im Falle des Rekanalisationsstatus oft einzelne Modelle je nach Rekanalisationsstatus trainiert wurden und andere Patientenfeatures nahezu ausschließlich als redundante Featurewerte in Bezug auf alle Voxel eines Patienten modelliert werden. Hierarchische Modellierungsprobleme wie diese, bei dem jeder Patient eine Vielzahl an Voxeln zu den Trainingsdaten beiträgt, unterliegen allerdings leicht einem systematischen Bias, da Patientenfeatures gegenüber anderen Features aus der Bildgebung unverhältnismäßig stark in die Prognose des Modells eingehen. Dies führt nicht nur dazu, dass die vorhandenen Informationen nicht optimal für das Lernen der Modelle ausgeschöpft werden, sondern dass es auch zu Überan-

passungen an die Trainingsdaten auf Basis der Patientenauswahl für die Modellierung kommen kann. Dies führt zu Schwierigkeiten in der Generalisierbarkeit der finalen Modelle.

Lokale Modelle sind bekannt dafür, dass sie bessere Ergebnisse für Teilmengen der Trainingsdaten liefern als globale Modelle [Hand und Vinciotti, 2003]. Sie wurden jedoch bisher nicht auf dem Gebiet der Schlaganfallprognose erprobt. Eine erfolgreiche lokale Modellierung für unterschiedliche Positionen im Gehirn würde es nicht nur erlauben, unterschiedliche anatomische und pathophysiologische Voraussetzungen für dieses Vorhersageproblem zu nutzen, sondern auch gleichzeitig eine konzeptionelle Lösung für die Integration von Patientenfeatures aufzeigen, da in jedes lokale Modell nur das Voxel der Modellposition eines Patienten in die Trainingsdaten eingeht. Somit findet keine erhöhte Gewichtung von Patientenfeatures gegenüber anderen Features aus der Bildgebung statt.

Aus diesem Grund wird in dieser Arbeit zunächst ein lokaler Ansatz entwickelt, bei dem exakt ein Modell pro Gehirnposition trainiert wird. Bei der reinen Form dieses Ansatzes gehen ausschließlich Voxel derselben Gehirnposition in ein lokales Modell ein, sodass räumliche Unterschiede durch die lokale Auswahl der jeweiligen Trainingsdaten modelliert werden. Zusätzlich erlaubt dieser Ansatz, Features auf Patientenebene zu integrieren, ohne dass es dadurch zu einer Verzerrung des Modells kommt, denn bei diesem Ansatz trägt jeder Patient nur ein Voxel zu jedem Modell bei. In zukünftigen Studien können Patientenfeatures wie der Rekanalisationsstatus als kategorielle Features in ein lokales Modell integriert werden, anstatt die Trainingsdaten wie z. B. bei [Winder et al., 2019] nach Rekanalisationskohorten aufsplitten zu müssen, was zu einer geringeren Trainingsdatenmenge führt. Vermieden wird dadurch auch, dass diese kategorielle Variable wie bei [Benzakoun et al., 2021] in überrepräsentierter Form ins Modell einfließt.

Eine Erweiterung dieses Ansatzes stellt der hybride Ansatz dar, dessen Voxel-Prädiktionen aus dem Mittel der entsprechenden lokalen Vorhersage und der eines globalen Modells bestehen. Dies vereint die Vorteile des lokalen Ansatzes mit denen eines globalen Modells: Während der Erste spezifische Trainingsdaten für die relevante Infarktposition liefert und damit eine Integrationsmöglichkeit von Patientenfeatures ohne Verzerrungen erlaubt, bietet das globale Modell die Anwendung auf wesentlich mehr Trainingsdaten und damit eine größere Evidenz.

Weiter wird in dieser Arbeit die Integration von räumlichen Features in eine globale Modellierung erforscht. Dazu werden zum ersten Mal die Läsionshäufigkeiten, die bereits im Tierversuch bei [Shen und Duong, 2008] und später beim Menschen von [Kemmling et al., 2015] im Rahmen einer logistischen Regression getestet wurden, in nichtlinearen Modellen wie Random Forest und XGBoost für die Vorhersage des Gewebeschicksals beim Schlaganfall eingesetzt. Insbesondere werden die in [Benzakoun et al., 2021] genutzten MNI-Koordinaten separat als räumliches Feature (ohne Nachbarschaftsmodellierung) validiert. Dazu wird der Einfluss der räumlichen Features in der logistischen Regression und in den tree-basierten Modellen validiert, und für die tree-basierten Modelle eine ausführliche Betrachtung der Features mittels Shapley-Werten vorgenommen, um Unterschiede und Gemeinsamkeiten von der Nutzung der räumlichen Features durch die tree-basierten Modelle festzustellen.

Im nächsten Kapitel werden diese Modellierungsansätze entwickelt, um sie dann zuerst in Bezug auf ihre Vorhersageergebnisse und den Effekt der räumlichen Prädiktoren zu kontrastieren. Es bleibt zu beantworten, welche Ansätze und Algorithmen von den räumlichen Informationen wie stark profitieren, und was es bei ihrer Integration zu beachten gilt. Weiterhin ist fraglich, ob die Datenmenge für die lokale und hybride Modellierung ausreicht, um eine aussagekräftige Verbesserung gegenüber einem globalen Modell ohne räumliche Informationen zu erzielen.

Als Grundlage für die Modellierung und die Beantwortung dieser Fragen dienen die Daten aus der multizentrischen I-KNOW-Studie, die im nächsten Abschnitt vorgestellt und danach spezifisch für die räumliche Modellierung in dieser Arbeit aufbereitet werden.

3 Vergleich zweier voxelbasierter Modellierungsansätze zur Integration räumlicher Informationen in die Vorhersage des Gewebeschicksals (lokal vs. global)

3.1 I-KNOW-Studie

Zur Modellierung der beiden Ansätze wurden MRT-Sequenzen und klinische Daten aus der I-KNOW-Studie[23] verwendet, in der MRT-Daten von Patienten mit Schlaganfällen im vorderen Stromgebiet an den Universitätskliniken der Städte Aarhus, Cambridge, Girona, Hamburg und Lyon von 2006 bis 2009 erhoben wurden. Das Ziel der I-KNOW-Studie war es, mithilfe dieser Daten die Entwicklung multivariater Modelle zur Vorhersage des Infarktrisikos auf Voxelebene zu ermöglichen. Neben der Genehmigung der Studie durch die lokalen Ethikkommissionen gaben die Studienteilnehmenden (oder ihre gesetzlichen Vertreter) ihre schriftliche informierte Einwilligung [Alawneh et al., 2011].

3.1.1 Patienten

Mit der vorliegenden Arbeit sollte insbesondere der Patientenkreis mit einem längeren Infarkt-Zeitfenster erforscht werden, bei der die medizinische Entscheidung über eine Rekanalisationstherapie basierend auf deren Erfolgsaussichten getroffen werden muss. Daher eignet sich das Sample der I-KNOW-Studie, weil dort hauptsächlich Patienten mit einem länger zurückliegenden Infarktbeginn berücksichtigt wurden.

Die Einschlusskriterien für die Patientenauswahl der I-KNOW-Studie waren (1) NIHSS ≥ 4, (2) eine abgeschlossene akute MRT-Untersuchung innerhalb von (a) sechs Stunden nach Infarktbeginn (im Falle einer Thrombolyse) oder (b) zwölf Stunden nach Infarktbeginn (im Falle einer konservativen Behandlung) und (3) ein durch DWI- und/oder PWI-Bildgebung nachweisbarer akuter Infarkt im vorderen Stromgebiet [Illig, 2016]**.**

Ausgeschlossen wurden Patienten mit (1) unklarem Zeitpunkt des Infarktbeginns, (2) sichtbarer Blutung in der T2*-gewichteten Bildgebung, (3) lakunärem Infarkt oder (4) einem Infarkt im hinteren Stromgebiet [Illig, 2016].

[23] Die Abkürzung im Projektnamen steht für “Integrating information from molecules to man: knowledge discovery accelerates drug development and personalized treatment in acute stroke” (https://cordis.europa.eu/project/id/027294/it).

Von jedem Patienten wurde sowohl eine Sozial- als auch eine Krankenanamnese in einer Datenbank aufgenommen: u. a. das Alter, das Geschlecht, der NIHSS, die Seite des Infarkts, die Zeit von Infarktbeginn bis zur Bildgebung, die medizinische Historie, vaskuläre Risikofaktoren, Medikamenteneinnahme seit Infarktbeginn, durchgeführte Behandlungen (vor allem Thrombolyse), die Ursache des Infarkts und das funktionelle Outcome (mittels mRS) [Alawneh et al., 2011].

Alle Patienten wurden bei ihrer Aufnahme mithilfe der MRT untersucht. Das Protokoll beinhaltete neben einer DWI-, einer FLAIR-, einer T2*-gewichteten Gradientenecho- und einer PWI-Sequenz auch eine Time-of-Flight-MR-Angiographie. Die gleichen Sequenzen wurden nach drei Stunden in der ersten FU-Untersuchung wiederholt. Weitere FU-Untersuchungen erfolgten nach zwei bis drei Tagen und nach einem Monat, wobei erneut dieselben Sequenzen mit Ausnahme der PWI erhoben wurden [Illig, 2016].

3.1.2 MRT-Protokoll

In die Modellierung dieser Arbeit sind in erster Linie die akuten DWI- und PWI-Datensätze eingeflossen, die innerhalb der ersten zwölf Stunden bei den Patienten nach dem Einsetzen der Symptome erhoben wurden, sowie die Follow-up-FLAIR-Sequenzen in einem Zeitraum von zwei bis drei Tagen nach Infarktanzeichen. Die Daten aus der weiteren Follow-up-Untersuchung, die einen Monat nach dem Infarkt stattfand, wurden nicht für die Modellierung herangezogen.

Bei den MRT-Untersuchungen wurden Magneten mit einer Feldstärke von 1,5 Tesla eingesetzt. Die einzelnen Sequenzen wurden mit den folgenden Parametern erhoben:

(1) Die DWI-Aufnahmen wurden mit Gradientenfeldstärken von b = 1000 s/mm^2, gemittelt für drei oder zwölf Richtungen, und b = 0 s/mm^2, TE = 100 ms und TR > 5 Sekunden durchgeführt. Die räumliche Auflösung innerhalb der einzelnen Schichten reichte bei dieser Sequenz von 0,9 × 0,9 mm^2 bis 2,0 × 2,0 mm^2. Die Schichtdicke betrug 3 oder 5 mm.

(2) Die FLAIR-Sequenzen wurden mit einer TE von 100 ms, einer TR von > 8 Sekunden und einer Inversionszeit von 2,5 Sekunden bei einem Flipwinkel von 150° aufgenommen. Die Auflösung innerhalb der Schichten betrug zwischen 0,45 × 0,45 mm^2 und 1,0 × 1,0 mm^2. Die Schichtdicke umfasste 5 mm. Der Abstand zwischen den 24 Schichten betrug je 1 mm. Die PWI wurde nach Gabe von 0,1 mmol/Kg Gadolinium

Kontrastmittel, gefolgt von 30 ml Kochsalzlösung mit einer TE von 30–50 ms und einer TR von 1,5 Sekunden, erhoben. Die räumliche Auflösung der PWI-Datensätze lag ebenfalls zwischen 0,9 × 0,9 mm^2 bis 2,0 × 2,0 mm^2, wobei die Schichtdicke der PWI-Aufnahmen 6 bis 6,5 mm umfasste.

Weitere Details zu den MRT-Protokollen der I-KNOW-Studie werden in den Arbeiten von [Cheng et al., 2014], [Ozenne et al., 2015] und [Berner et al., 2016] angegeben.

3.2 Berechnung der Bildfeatures für die Modellierung

Die unmittelbar bei der Patientenaufnahme im Krankenhaus erhobenen DWI-, PWI- und FLAIR-Sequenzen und die zwei bis drei Tage später erhobenen Follow-up-FLAIR wurden zunächst mittels der Software AnToNIa prozessiert. Dieses Programm wurde am Institut für Medizinische Informatik des Universitätsklinikums Hamburg-Eppendorf (UKE) in Zusammenarbeit mit der Arbeitsgruppe für klinische Schlaganfallbildgebung entwickelt [Forkert et al., 2009]. Das Akronym AnToNIa steht für Analysis Tool for Neuro Imaging Data. Mit der ersten Version gelang es, Fragestellungen hinsichtlich der Darstellung zerebraler arteriovenöser Malformationen zu beantworten, die eine wichtige Ursache für Gehirnblutungen sind. Die dieser Arbeit zugrunde liegende Version „Perfusion and Stroke“ [Forkert et al., 2014] wurde speziell zur Evaluierung von multimodalen Schlaganfalldaten entwickelt. Damit ermöglicht sie kombinierte Auswertungen von DWI-, PWI- und Follow-up-FLAIR-Sequenzen, sodass mehrere Parameter gleichzeitig in die Modellierung der beiden Ansätze aufgenommen werden können. Im Folgenden wird das Prozessieren der Datensätze für eine 43-jährige Patientin aus der I-KNOW-Studie illustriert.

Berechnung der Parameterkarten

Als Erstes wurde aus den DWI-Sequenzen der ADC berechnet. Auf Basis eines Schwellwerts von $1200 \cdot 10^{-6}$ mm^2/s wurde das Gehirn in Gewebe (niedriger ADC) und Liquor cerebrospinalis (CSF, engl. cerebrospinal fluid; hoher ADC) unterteilt [Forkert et al., 2014]. Zur Unterscheidung der Gehirnhälften wurden zwei Linien in verschiedenen Gehirnschichten definiert. Diese dienen als Approximation der hemisphärischen Fissur, sodass sich das Gehirn demnach in eine ipsi- und kontralaterale Hemisphäre gliedern lässt. In Abbildung 8 sind die Aufnahmen aus der DWI und der daraus be-

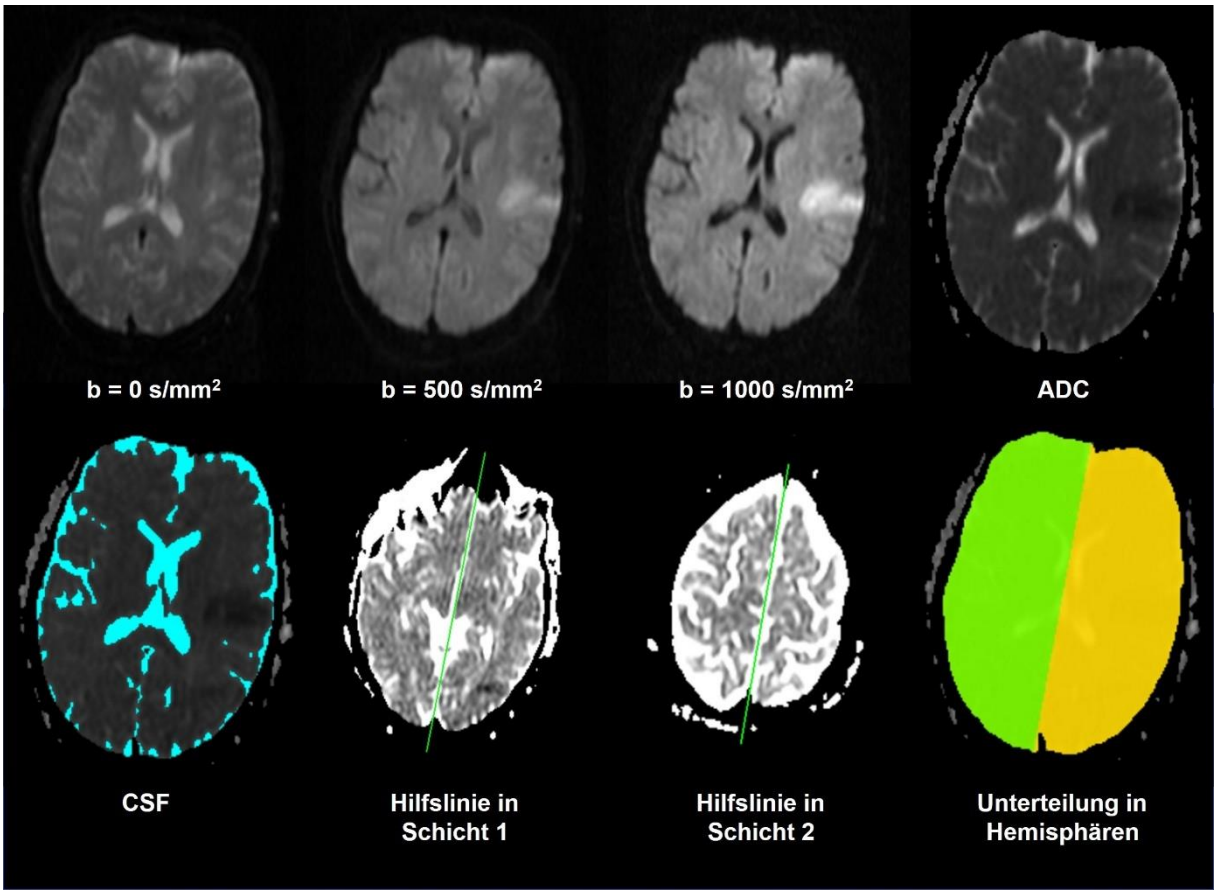

Abbildung 8: Berechnung des ADC, Segmentierung des CSF und Unterteilung des Gehirns für eine ausgewählte Patientin (43 Jahre, NIHSS = 6): Obere Reihe: DWI-Aufnahmen und der daraus resultierende ADC. Untere Reihe: Segmentierung des CSFs (hoher ADC) und manuell definierte Hilfslinien in zwei weit voneinander entfernten Schichten zur Unterteilung der beiden Gehirnhälften (rechts in grün, links in dunkelgelb) (Quelle: I-KNOW-Daten).

rechnete ADC sowie die Segmentierung des CSFs und die Einteilung der Gehirnhälften auf Basis der manuell definierten Hilfslinien dargestellt.

Zur Berechnung der Perfusions-Parameter wurden die PWI-Sequenzen als Erstes um Bewegungsartefakte korrigiert (engl. in-slice motion correction). Hierzu wurden die einzelnen Aufnahmen auf die jeweils vorherige Aufnahme registriert, sodass alle Aufnahmen sukzessive auf die initiale PWI-Aufnahme verschoben werden. Das Anfluten des Kontrastmittels, dessen erste Passage durch das Gefäßbett und das Abfließen lassen sich anhand der sequenziellen PWI-Aufnahmen für die einzelnen Schichten nachvollziehen (siehe Abbildung 9). Hier zeigt der Vergleich zwischen den einzelnen Gehirnhälften schon eine Verspätung bzw. ein Ausbleiben des Kontrastmittels in Teilen der linken Gehirnhälfte der Patientin und damit ein akutes Durchblutungsdefizit an.

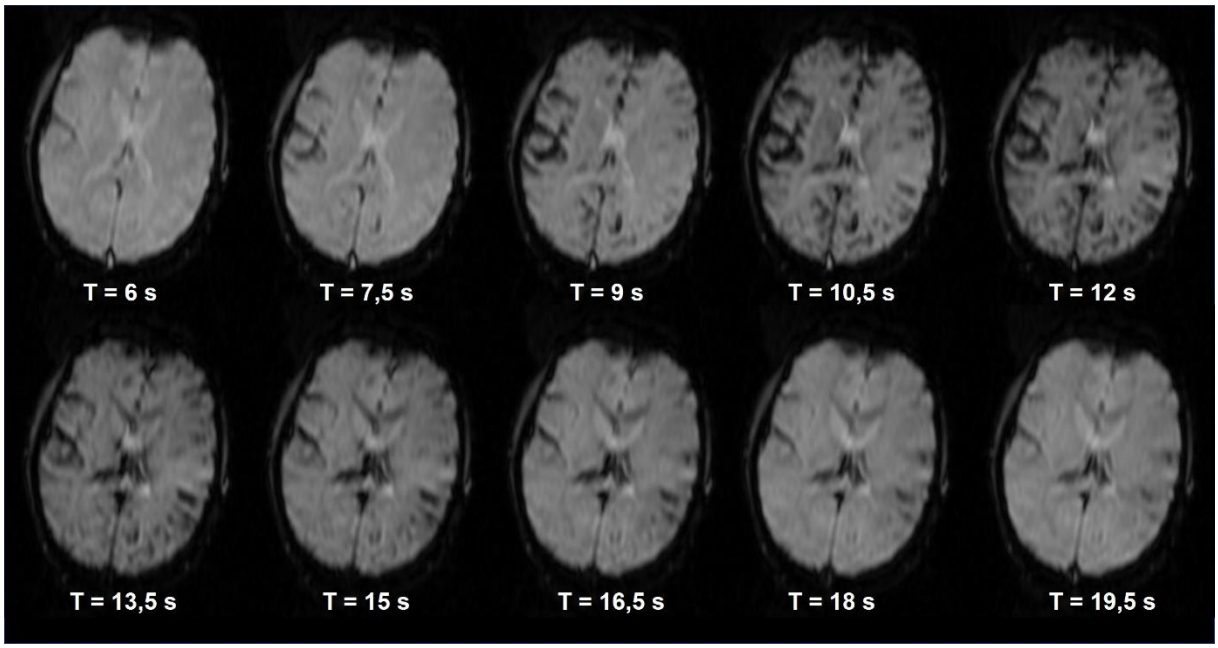

Abbildung 9: PWI-Sequenz nach Korrektur von Patientenbewegungen für verschiedene Zeitpunkte bei einer ausgewählten Patientin (43 Jahre, NIHSS = 6): PWI-Aufnahmen einer Gehirnschicht nach Injektion des Kontrastmittels. Während zu Beginn das Kontrastmittel anflutet, sind danach die Ankunft und der Verlauf des Kontrastmittels gut zu erkennen (Quelle: I-KNOW-Daten).

Da benachbarte Schichten nicht gleichzeitig gemessen werden können und die Messpunkte um bis zu TR/2 (= 0,75 s) voneinander abweichen, kann es passieren, dass die Daten die zeitliche Reihenfolge des Blutflusses durch die einzelnen Schichten nicht korrekt repräsentieren, was zukünftige Analysen negativ beeinflusst. Da jedoch die Messzeitpunkte für die einzelnen Schichten bekannt sind, können für jedes Voxel geglättete Konzentrations-Zeit-Kurven berechnet werden, sodass daraus approximierte Kontrastmittelkonzentrationen zu einheitlichen Zeitpunkten für alle Voxel gewählt werden können und der Blutfluss zwischen den einzelnen Schichten korrekt abgebildet wird. In AnToNIa wurde daher eine temporale Interpolation mit Zeitabständen von einer Sekunde für die Kontrastmittelwerte vorgenommen. Zur Anpassung der Konzentrations-Zeit-Kurven wurde die B-Spline-Methode eingesetzt [Forkert et al., 2014].

Für die dekonvolutionsbasierte Berechnung von PWI-Parameterkarten muss die AIF, d. h. die Konzentrations-Zeit-Kurve für das Anfluten des Kontrastmittels in die Zielregion, bestimmt werden. Zur Eingrenzung des Gebiets, das für die Lokalisation der AIF infrage kommt, wurde zunächst eine Gefäßkarte, die die typische Lokalisationen der Hals- und mittleren Gehirnschlagader enthält [Forkert et al., 2013], auf die Perfusionsaufnahmen abgebildet. Hierzu wurde zuerst der MNI-Atlas affin auf ein

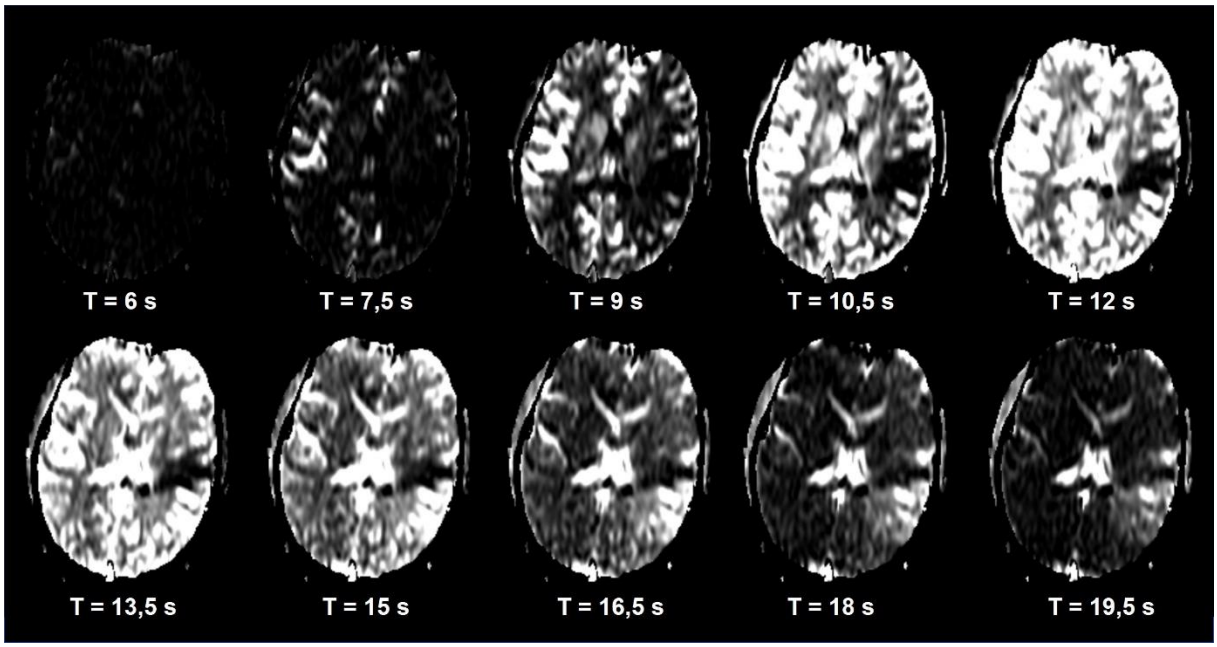

Abbildung 10: PWI-Sequenz nach Anwendung der Dekonvolution für verschiedene Zeitpunkte bei einer ausgewählten Patientin (43 Jahre, NIHSS = 6). Für einen stärkeren Kontrast wurde eine Aufnahme ohne Kontrastmittel abgezogen (Quelle: I-KNOW-Daten).

Durchschnitts-PWI-Bild ohne Kontrastmittel registriert. Die daraus erhaltene Transformation wurde genutzt, um die Gefäßkarte, die auf den Atlasinformationen basiert, kozuregistrieren. Mit den Gefäßinformationen wurde nun eine Segmentierung (inkl. eines kleinen Puffers) der relevanten Lokalisationen für die AIF-Auswahl vorgenommen. Um arterielle und nicht-arterielle Signale innerhalb dieser Segmentierung unterscheiden zu können, wurden die dortigen Konzentrationszeitkurven mithilfe eines k-Means-Clustering Ansatzes unterteilt. Danach wurden die Kurven des Clusters mit frühen und hohen Time-to-Peaks, die auf eine Arterie hinweisen, mithilfe einer geometrisch korrekten Methode zu einer finalen AIF gemittelt [Forkert et al., 2014]. Hiermit konnten die Konzentrations-Zeit-Kurven aller Gewebevoxel um die AIF korrigiert werden (Abbildung 10). Für die Berechnung der Dekonvolution wurde dabei eine blockzyklische Singulärwertzerlegung (engl. block-circulant singular value decomposition) mit einem Schwellwert von 0,15 verwendet, die in [Wu et al., 2003] näher beschrieben ist.

Aus den korrigierten Konzentrations-Zeit-Kurven wurden dann die Perfusionsparameterkarten berechnet (CBF, CBV, MTT und T_{max}). Nach einer rigiden Transformation auf das Baseline-DWI-Bild ($b = 0$ s/mm^2) der Patienten wurden die PWI-Parameter an der gesunden Gehirnhälfte normalisiert, um diesbezügliche Variationen zwischen den

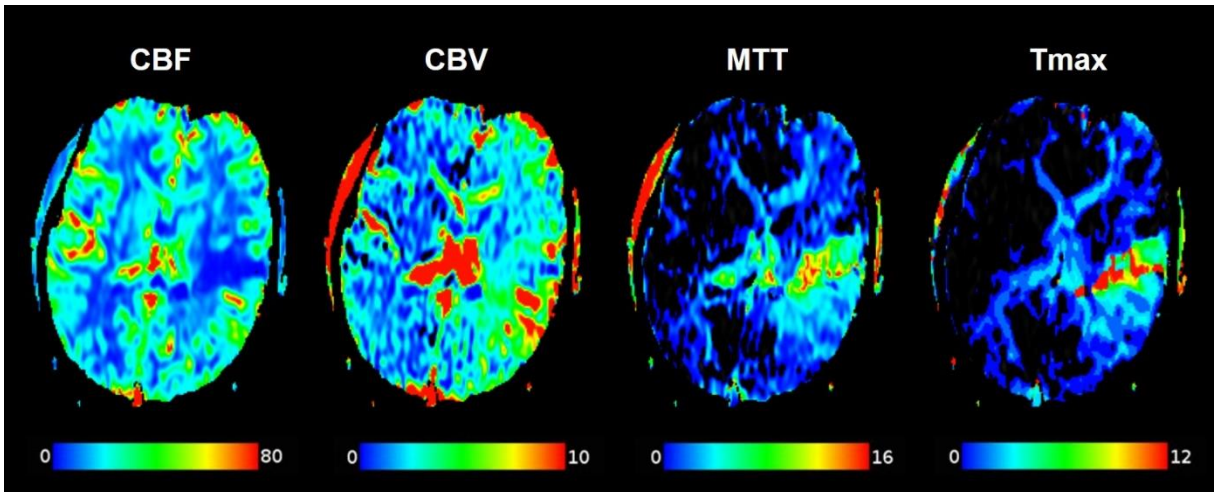

Abbildung 11: Perfusionsparameter nach Dekonvolution und Normalisierung mittels kontralateraler Hemisphäre für eine ausgewählte Patientin (43 Jahre, NIHSS = 6). Normalisierte CBF- (von 0–80 ml/min/100 g), CBV- (von 0–10 ml / 100 g), MTT- (von 0–16 Sekunden) und T_{max}- (von 0–12 Sekunden) Karten (Quelle: I-KNOW-Daten).

Patienten auszugleichen: Für die beiden Zeitparameter MTT und T_{max} wurde der Durchschnittswert der kontralateralen Seite (ohne CSF) subtrahiert. Die Kontrastmittelverzögerungen können durch verschiedene Aspekte wie das Injektionsprotokoll und die Herzzeitvolumenfunktion beeinflusst sein, sodass die Differenz zwischen der betroffenen und der nicht betroffenen Seite hier informativer als ein relativer Wert ist. Für CBF und CBV wurden die Werte der ipsilateralen Hemisphäre durch den Durchschnittswert der kontralateralen Seite geteilt. Für die CBF- und CBV-Karten ist die Magnitude der Messwerte eine wichtige Eigenschaft des Gewebes, sodass deren relative Werte zur nicht betroffenen Hemisphäre bedeutsamer sind als eine absolute Differenz. Siehe Abbildung 11 für eine Darstellung der finalen PWI-Parameter der 43-jähringen Patientin aus der I-KNOW-Studie.

Segmentierung der Läsionsmasken

Die Infarktläsionen in den Follow-up-FLAIR-Datensätzen wurden von zwei Neuroradiologen jeweils manuell segmentiert (siehe Abbildung 12). Damit resultierten für jeden Patienten der ADC und die normalisierten PWI-Parameter sowie eine zugehörige Follow-up-FLAIR mit Läsionsmaske. Final wurden diese Karten ebenfalls auf die DWI ($b = 0$ s/mm^2) registriert, um für Patientenbewegungen zwischen den Aufnahmen und/oder verschiedene Blickfelder zu korrigieren.

Die Zwischenergebnisse wurden nach jedem Bearbeitungsschritt visuell überprüft. Ausgeschlossen wurden für die weitere Analyse dann diejenigen Datensätze mit feh-

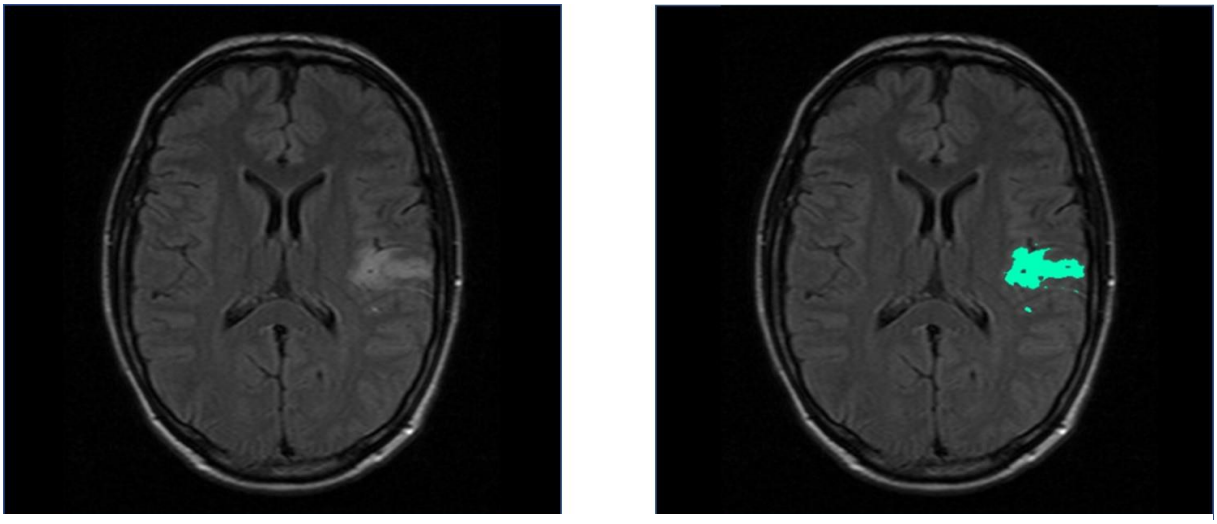

Abbildung 12: Follow-up-FLAIR-Sequenz und Läsionsmaske für eine ausgewählte Patientin (43 Jahre, NIHSS = 6): Follow-up-FLAIR ohne (links) und mit markierter Infarkt-Läsion (rechts) (Quelle: I-KNOW-Daten).

lenden klinischen oder für die Analyse benötigten Bildgebungsdaten, schlechter Bildqualität, ungenügender Kontrastmittelsichtbarkeit in der PWI-Sequenz, Artefakten in den Bilddaten oder nicht sichtbaren bzw. unzureichenden Folgeläsionen (siehe Abbildung 13).

Für die Berechnung der räumlichen Features in dieser Arbeit muss die Position eines Voxels für alle Patienten vergleichbar sein. Daher werden die Daten im nächsten Abschnitt auf einen Gehirnatlas registriert, dessen Koordinaten hierfür als Referenz dienen.

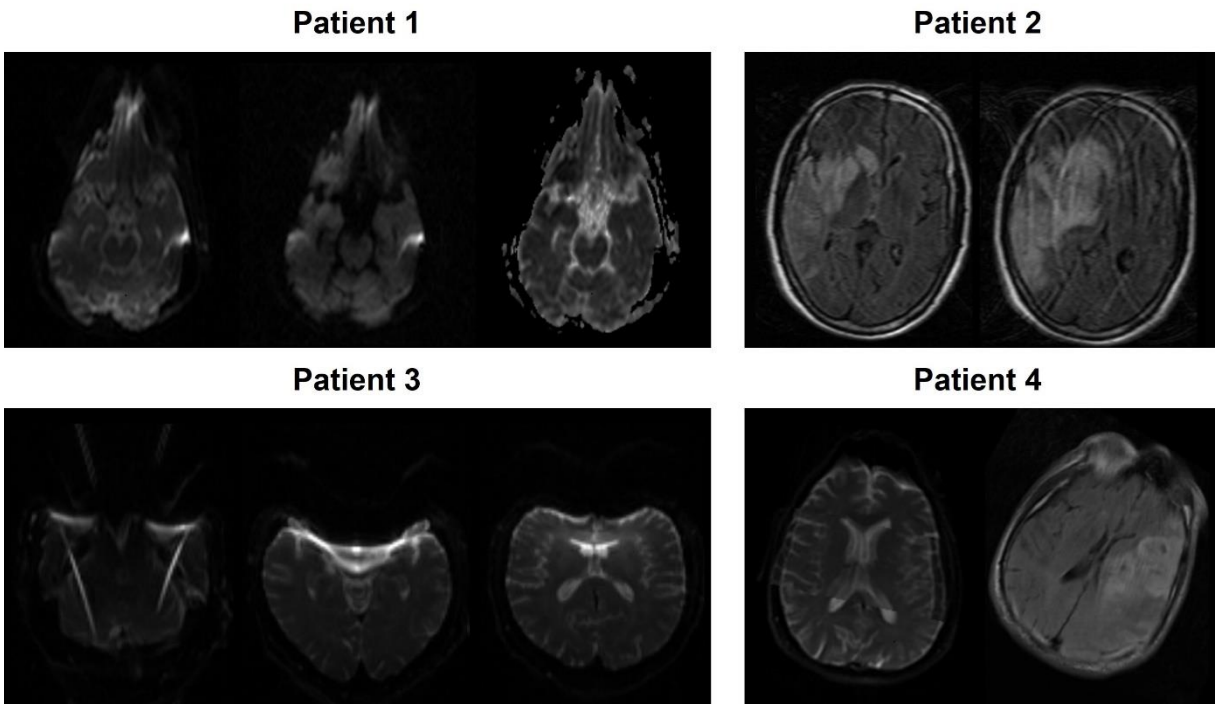

Abbildung 13: Beispiele für ausgeschlossene Patientendatensätze: Patient 1: Folding-Artefakte in der DWI für b = 0 s/mm^2 (links) und b = 1000 s/mm^2 (Mitte), die sich in der ADC-Karte (rechts) widerspiegeln. Patient 2: starke Bewegungsartefakte u. a. in der Baseline-FLAIR- (links) und FU-FLAIR-Sequenz (rechts). Patient 3: Rotationsartefakte in der DWI für b = 0 s/mm^2 in drei verschiedenen Schichten. Patient 4: leichte Artefakte in der DWI für b = 0 s/mm^2 (links) und fehlerhafte Registrierung (Quelle: I-KNOW-Daten).

3.3 MNI-Registrierung und Datenaufbereitung im MNI-Raum

Um Voxelpositionen zwischen den unterschiedlichen Patienten trotz unterschiedlicher Anatomie vergleichen zu können, wurden die Bild-Karten aus AnToNIa in das Statistikprogramm R (Version 3.4.2) geladen und mit dem ANTsR-Paket [Avants et al., 2017] nichtlinear auf eine repräsentative Abbildung des Gehirns, den (symmetrischen) MNI-Atlas, registriert.[24] Hierzu wurde der MNI-Atlas ICBM 2009c Nonlinear Symmetric mit einer Auflösung von 1 × 1 × 1 mm^3 unter http://nist.mni.mcgill.ca/icbm-152-nonlinear-atlases-2009/ ausgewählt. Zunächst wurden die DWI-Bilder (b = 0 s/mm^2) auf den MNI registriert, um danach die hieraus resultierende Transformation auf die anderen Karten anzuwenden, die zuvor auf die DWI registriert wurden. Um eine Verdopplung von Interpolationsfehlern zu vermeiden, fand eine Verkettung der Transformationen statt. Um für Änderungen der CSF-Verteilung durch Schwellungen zu korrigieren, wurden alle CSF-Voxel aus den registrierten Parameterkarten ausgeschlossen. Nach diesem

[24] Als Algorithmus wurde SyN mit linearer Interpolation für kontinuierliche Parameterkarten und mit Nearest-Neighbor-Interpolation für binäre Parameterkarten verwendet.

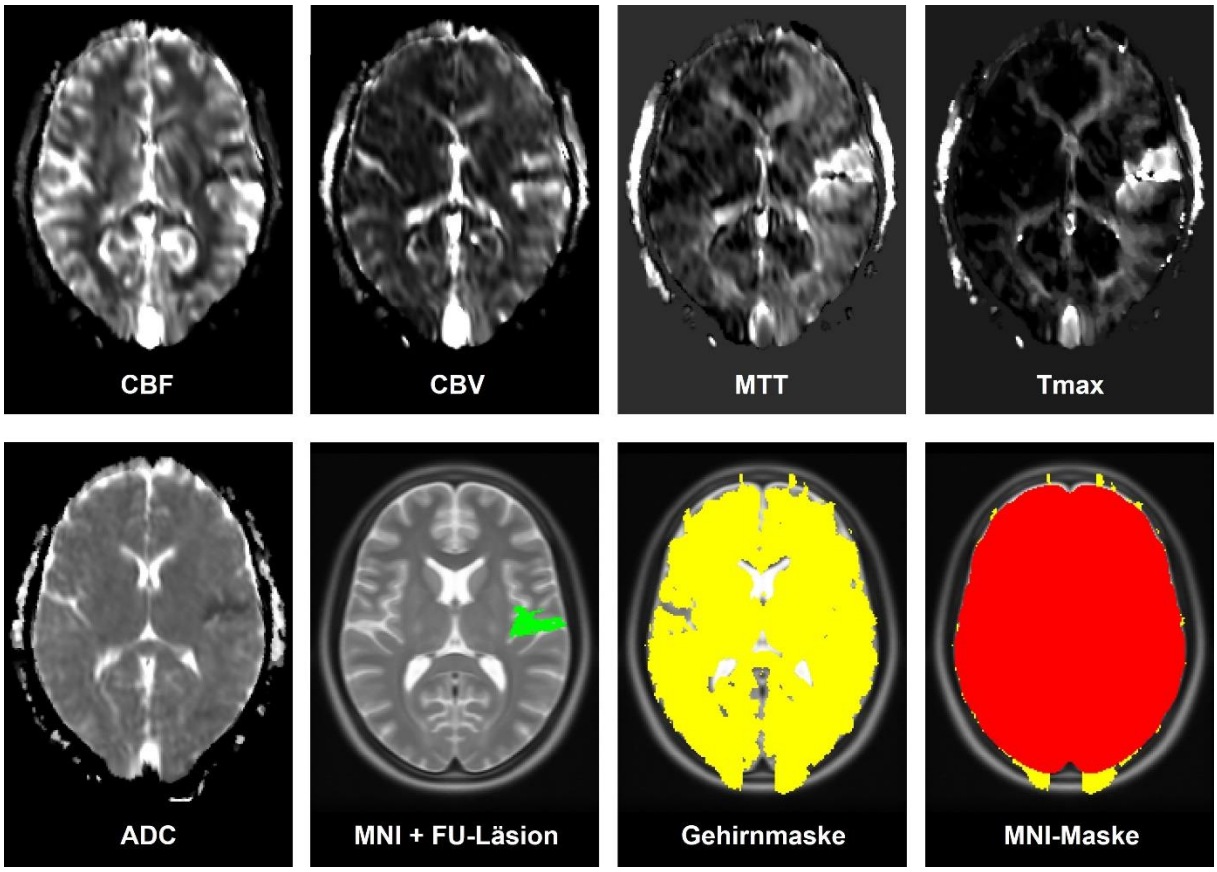

Abbildung 14: Parameterkarten und Segmentierungen im MNI-Raum für eine ausgewählte Patientin (43 Jahre, NIHSS = 6): Obere Reihe: registrierte CBF-, CBV-, MTT und T_{max}-Karten. Untere Reihe: registrierter ADC, MNI-Karte mit registrierter FU-Läsions-Maske (grün) und die registrierte Gehirnmaske der Patientin (ohne CSF; gelb) sowie die Gehirnmaske für den MNI-Atlas (rot). Sowohl die Gehirnmaske der Patienten als auch diejenige des MNI-Atlas wurden auf die Registrierungen der Parameterkarten zum Filtern der Voxel für die Modellierung angewandt (Quelle: I-KNOW-Daten).

Schritt liegen alle in dieser Arbeit verwendeten Sequenzen im MNI-Koordinatenraum, wobei visuell kontrolliert wurde, dass anatomisch auffällige Regionen wie z. B. die Ventrikel nach dem Morphen räumlich mit denen des MNI-Atlas übereinstimmen. Zusätzlich wurden die registrierten Sequenzen auf die Gehirnmaske des MNI gefiltert. Abschließend standen für jedes Voxel der ADC, die normalisierten PWI-Parameter (CBF, CBV, MTT, T_{max}), die MNI-Koordinaten (x, y und z-Achse) sowie die FU-Läsion mit dem jeweiligen binären Gewebeoutcome (0 = Nichtläsion und 1 = Läsion) zur Verfügung (siehe Abbildung 14).

Da es sich bei den Schlaganfällen im I-KNOW-Datensatz, wie bei der großen Mehrheit aller Schlaganfälle, ausschließlich um einseitige Infarkte handelt,[25] wurden alle Voxel der jeweiligen kontralateralen Gehirnhälfte ausgefiltert. Für die Datenverarbei-

[25] Etwa 9,4% der ischämischen Schlaganfälle sind bilaterale Infarkte [Michel et al., 2010].

tung im weiteren Teil dieser Arbeit wurden vor allem das arrayhelpers- [Beleites, 2016], dplyr- [Wickham et al., 2020] und data.table- [Dowle und Srinivasan, 2020] sowie das foreach-Paket verwendet [Revolution Analytics und Weston, 2015]. Analytische Grafiken wurden mit dem ggplot2-Paket [Wickham, 2016] und Heatmaps der Gehirnkarten im MNI-Raum mit dem FSL image viewer [Jenkinson et al., 2012] erstellt.

Nach den Vorarbeiten zur räumlichen Vergleichbarkeit der Gehirn-Karten wurden die Daten in die jeweilige Modellierung eingespeist. Zunächst wird der lokale Ansatz beschrieben, bei dem ein Modell pro MNI-Position trainiert wird.

3.4 Lokaler Ansatz

Der lokale Ansatz basiert auf der Idee, das Gewebeoutcome beim ischämischen Schlaganfall durch ein lokales Modell pro Position auf dem MNI-Atlas zu modellieren (siehe Abbildung 15). Dadurch wird jedes lokale Modell ausschließlich auf Voxeln mit dort üblichen Gewebetypen und den dort vorkommenden anatomischen Durchblutungsvoraussetzungen trainiert. Bei der Anwendung des lokalen Ansatzes zur Gewebeprognose auf einem neuen Patientendatensatz werden die Vorhersagen für jede Position von dem dort trainierten lokalen Modell getroffen, sodass anstelle einer allgemeinen Risikobewertung eine lagebedinge Prognose getroffen wird. Für jede Position der betroffenen Gehirnhälfte ist danach eine Prädiktion vorhanden, wodurch eine Karte des Risikos für Läsionen in der Follow-up-Untersuchung entsteht.

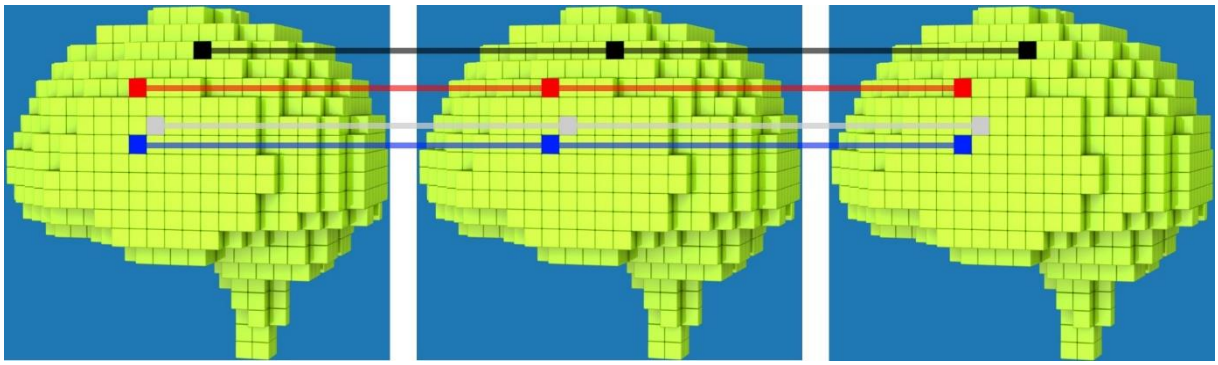

Abbildung 15: Grundidee des lokalen Ansatzes: Die Grundidee des lokalen Ansatzes besteht darin, ein lokales Modell pro MNI-Position zu trainieren. In diesen drei Abbildungen sind die Trainingsdaten von drei verschiedenen Patienten beispielhaft für vier lokale Modelle farblich dargestellt (schwarz, rot, grau und blau). Farblich gekennzeichnete Trainingsvoxel eines lokalen Modells sind dabei zusätzlich durch eine Gerade miteinander verbunden (Quelle: in Anlehnung an [Mirexon, o. J.], leicht modifiziert).

Da nach der Grundidee dieses Ansatzes nur maximal ein Voxel pro Position von jedem Patienten zum Training zur Verfügung steht, wurde der Ansatz mit zwei Algorithmen durchlaufen, die nicht nur schnell, sondern auch datensparsam sind: logistische Regression und Random Forest. Zum Training der einzelnen Modelle wurden der ADC und die vier normalisierten Perfusionsparameter CBF, CBV, MTT und T_{max} als Features und die FU-Läsionsmasken als Target verwendet. Sowohl das Training der lokalen Modelle als auch deren Anwendung zur Vorhersage wurden parallel durchgeführt. Anschließend wurden die Koeffizienten für die lokale logistischen Regression auf Unterschiede zwischen einzelnen funktionalen und anatomischen Gehirnregionen untersucht, um festzustellen, ob sie sich zwischen diesen Gebieten unterscheiden.

3.4.1 Anreicherung der lokalen Trainingsdaten

Für das Training eines robusten Machine-Learning-Modells braucht es eine heterogene Datenbasis, weil damit eine große Bandbreite an realistischen Fällen ins Modell eingespeist werden kann. So verfügt das Modell über eine breite Evidenzbasis, sodass eine Übertragbarkeit auf neue akute Patientendatensätze in der klinischen Praxis möglich ist. Da die I-KNOW-Studie Patienten mit Verschlüssen im vorderen Stromgebiet enthält, kommen naturgemäß nicht alle Gehirnpositionen für eine lokale Modellierung anhand dieses Datensatzes in Frage. Aufgrund der ungleichmäßigen räumlichen Verteilung der restlichen Läsionen in den Trainingsdaten gab es allerdings weitere Positionen im MNI, an denen kaum eine Läsion in den I-KNOW-Daten vorkam. Dadurch konnte dort kein lokales Modell trainiert werden. Dies ist auch in Abbildung 16, in der die heterogene Verteilung der Läsionen innerhalb der I-IKOW-Studie visualisiert ist, zu sehen. Deshalb wurde die Menge der Trainingsvoxel für jedes lokale Modell mittels zwei Anreicherungsdesigns schrittweise erweitert, indem pro Patienten jeweils zwei Voxel unterschiedlicher Outcomeklassen in die Trainingsdaten eines lokalen Modells im Idealfall eingehen (vgl. Abschnitt 4.2).

Zunächst bestehen die Trainingsdaten eines lokalen Modells aus allen Voxeln, die an der Modellposition im MNI-Raum liegen. Da hier keine CSF-Voxel und Voxel der kontralateralen Seite eingehen, weil diese bereits während der Datenverarbeitung aus dem Datensatz herausgefiltert wurden, entspricht die durchschnittliche Anzahl an Trainingsvoxeln eines lokalen Modells initial etwas weniger als der Hälfte der Patientendatensätze.

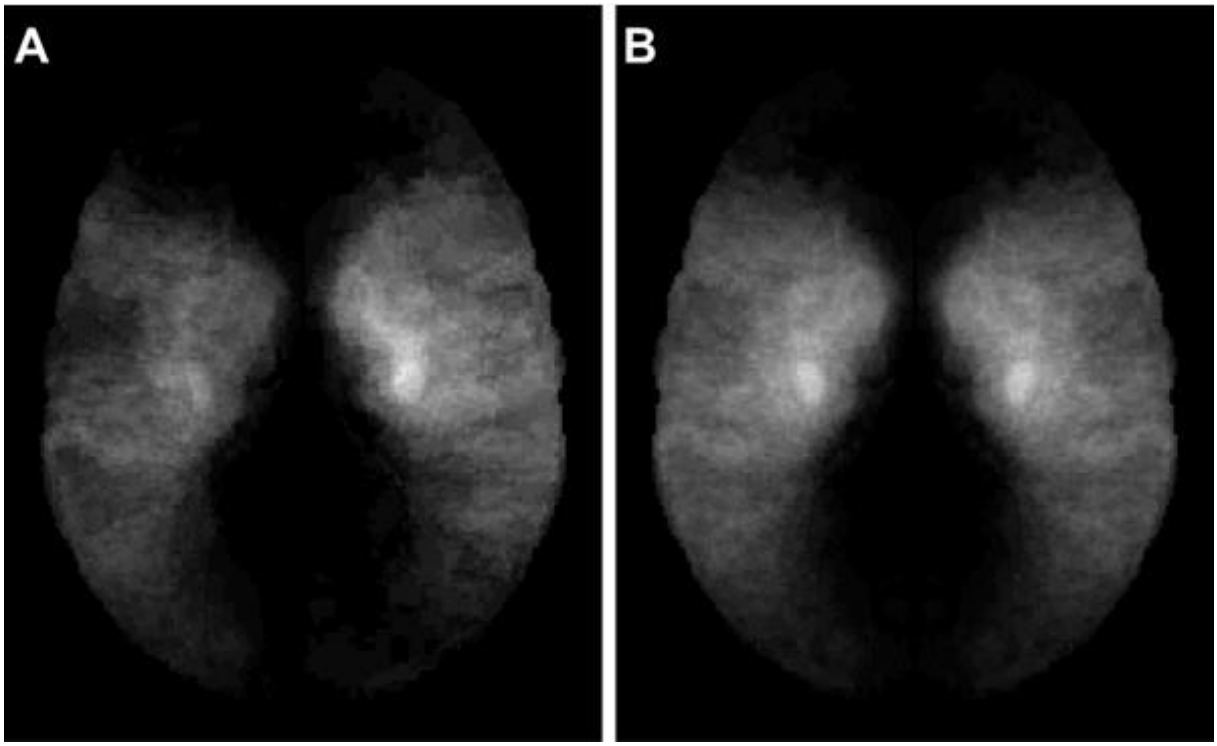

Abbildung 16: Räumliche Läsionsverteilung der I-KNOW-Patienten im MNI-Raum: Wahrscheinlichkeitsverteilung adjustiert nach Gehirnhälfte (A) vs. Wahrscheinlichkeitsverteilung basierend auf der in dieser Arbeit getroffenen Symmetrieannahme (B) (Quelle: I-KNOW-Daten).

Um dafür zu sorgen, dass Voxel beider Outcomeklassen in den Trainingsdaten vorkommen, wurden im ersten Anreicherungsschritt (maximal) noch zwei Voxel unterschiedlicher Outcomes aus der unmittelbaren Umgebung der Modellpositionen pro Patientendatensatz hinzugefügt. Zur Auswahl dieser Voxel wurde um jede Modellposition eine Kugel betrachtet, deren empirisch ausgewählter Kugelradius mit $\sqrt{5}$ Voxeln bewusst geringgehalten wurde. So konnte das Kriterium dieses Ansatzes, die Lokalität der angereicherten Voxel, erfüllt werden. Praktisch wurde der Anreicherungsradius über den euklidischen Abstand implementiert: Für ein Voxel an der Modellposition x_0, y_0, z_0 wurde der Abstand zu einem Voxel an der Position x_1, y_1, z_1 als

$$d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

berechnet. Ignoriert wurden bei der Anreicherung diejenigen Voxel, die außerhalb dieser Kugel liegen. Innerhalb der Kugel wurden der jeweils nächstgelegene Läsions- und Nichtläsionsvoxel des Patienten ausgewählt. Falls ein Patient an der Modellposition kein Gewebevoxel aufwies, wurden (maximal) beide Voxel angereichert – ansonsten (maximal) einer. So trug nach diesem Schritt jeder Patient maximal zwei Trainingsvoxel zu jedem der lokalen Modelle auf seiner Infarktseite bei. Durch diesen Anreiche-

rungsschritt konnten insgesamt zwar weitaus mehr lokale Modelle trainiert werden als zuvor, dennoch gab es weiterhin viele Voxelpositionen im MNI-Atlas, an denen eine zu geringe Anzahl an den insgesamt selteneren Läsionsvoxeln in den Trainingsdaten vorkam, um ein robustes Modell zu trainieren (vgl. Abschnitt 4.2).

Aus diesem Grund wurde ein zweiter Anreicherungsschritt vorgenommen. Nachdem im ersten Schritt nur Voxel der Gehirnhälfte der Modellposition in die lokalen Trainingsdaten einbezogen wurden, enthielten die jeweiligen Trainingsdaten keine Voxel von Patienten mit Infarkten auf der jeweils anderen Gehirnhälfte. Da von einer bilateralen morphologischen Symmetrie des Gehirns auszugehen ist, wurden nun für alle Modelle die Trainingsdaten der jeweils gegenüberliegenden Modelle (symmetrisch zur Hyperebene bei x = 0 des symmetrischen MNI-Raums) einbezogen. Mit anderen Worten, es wurden die Trainingsdaten der lokalen Modelle unabhängig von ihrer Gehirnhälfte in jeweils einem Trainingsdatensatz vereint. Nach diesem Schritt trägt ein Patient zwar weiterhin maximal zwei Voxel zu den Trainingsdaten eines einzelnen lokalen Modells bei, aber die Anzahl der Patienten (und damit der Trainingsdaten) verdoppelt sich (vgl. Abschnitt 4.2).

Nach diesen beiden Anreicherungsschritten konnten zwar an deutlich mehr MNI-Positionen lokale Modelle trainiert werden, dennoch kamen wesentlich weniger Läsionsvoxel als Nichtläsionsvoxel in den lokalen Trainingsdatensätzen vor, weil Nichtläsionen im Schlaganfallgehirn durchschnittlich überwiegen. Um sicherzustellen, dass die einzelnen Trainingsdatensätze Informationen zu beiden Outcomeklassen beinhalteten, wurde ein Schwellwert von mindestens drei Voxeln beider Klassen in den Trainingsdaten vorausgesetzt, um ein lokales Modell zu trainieren. Dieser Schwellwert erlaubte es weiterhin, an vielen Positionen mit niedrigem Infarktvorkommen ein lokales Modell zu trainieren, sodass der lokale Ansatz für einen großen Anteil an MNI-Positionen angewendet werden konnte. Dennoch fehlten nach diesem Schritt Modelle für lokale Trainingsdatensätze mit weniger als drei Läsionsvoxeln, denn Nichtläsionsvoxel waren stets genügend vorhanden. Deshalb wurde der lokale Ansatz ausgeweitet (vgl. Abschnitt 3.4.2).

3.4.2 Erweiterung des lokalen Ansatzes

Um trotz der fehlenden lokalen Modelle eine Prognose auf einer gesamten Gehirnhälfte (inkl. dem hinteren Stromgebiet) mit dem lokalen Ansatz treffen zu können, wurden auf einer Zufallsauswahl der gesamten Datenbasis eine logistische Regression

und ein Random trainiert. Da die Trainingsdaten und der Anwendungsbereich dieser Modelle den gesamten MNI-Raum abdecken, werden der Ansatz und die Modelle in dieser Arbeit als globaler Ansatz bzw. globale Modellierung bezeichnet (vgl. Abschnitt 3.4.3). Mit dem globalen Ansatz werden die Lücken im lokalen Ansatz gefüllt, weil dadurch Vorhersagen an Positionen, an denen kein lokales Modell trainiert werden konnte, durch die entsprechenden Vorhersagen des globalen Modells (je nach verwendetem Algorithmus, logistische Regression oder Random Forest) substituiert werden können.

3.4.3 Globaler und hybrider Ansatz

Der globale Ansatz ist dennoch als eigenständiger Ansatz zu verstehen, denn die einzelnen globalen Modelle (logistische Regression und Random Forest) wurden analog zum lokalen Ansatz nach derselben Methodik validiert. Somit konnte die Prognosegüte des lokalen Ansatzes mit der des (gängigen) globalen Ansatzes verglichen werden. Im Gegensatz zu einem einzelnen lokalen Modell fällt das Training der globalen Modelle aufgrund der hohen Voxelanzahl im MNI-Raum rechenintensiver aus, denn hierfür wurden zufällig 500.000 Voxel als Trainingsdaten gezogen, die nach Outcome stratifiziert sind. Damit lässt sich das Verhältnis aus Nichtläsions- und Läsionsvoxeln kontrollieren, sodass dieses möglichst ähnlich zu den Verhältnissen in den lokalen Trainingsdatensätzen war und der globale Ansatz besser den lokalen Ansatz vervollständigen konnte.

Zusätzlich wurde in dieser Arbeit noch ein hybrider Ansatz gewählt, der dem Mittelwert aus lokalem und globalem Ansatz entspricht, und damit zwar weiterhin die lokal unterschiedlichen anatomischen Voraussetzungen berücksichtigt, jedoch auf wesentlich mehr Evidenz bzgl. des ADC und der PWI-Parameter basiert. Außerdem glättet er die Prädiktionen des lokalen Ansatzes, die auch aufgrund der unterschiedlichen Anzahlen von Läsionsvoxeln in den Trainingsdaten stark variieren können. Falls kein lokales Modell für eine Position trainiert wurde, entsprach die dortige Vorhersage des hybriden Modells derjenigen des globalen Modells für den jeweils verwendeten Algorithmus (logistische Regression oder Random Forest). Der hybride Ansatz wurde nach derselben Methodik und anhand derselben Metriken wie der lokale und der globale Ansatz validiert.

Aus der Kombination der drei Ansätze (lokal, global und hybrid) und den zwei eingesetzten Algorithmen (logistische Regression und Random Forest) ergeben sich ins-

gesamt sechs Modelle. Jedes der Modelle wurde anhand des ADC und der normalisierten Perfusionsparameter (CBF, CBV, MTT, T_{max}) darauf trainiert, das (im Training) bekannte Gewebeoutcome (Läsion oder Nichtläsion) vorherzusagen.

3.5 Globaler Ansatz mit räumlichen Features

Der globale Ansatz entspricht dem gängigen Ansatz zur Vorhersage des Gewebeoutcomes. Er wurde hier zur Erweiterung des lokalen Ansatzes, der Entwicklung des hybriden Ansatzes und zum Vergleich mit ihnen eingeführt. In diesem Abschnitt wird der globale Ansatz um Features erweitert, welche die räumlichen Informationen explizit abbilden, die in den lokalen Modellen implizit verwendet wurden. Im Vergleich zum hybriden Ansatz bietet solch eine Modellierung den Vorteil, dass das Modell nicht aus einem lokalen und einem globalen Modell zusammengesetzt ist, sondern gleichzeitig die Lageinformationen aller Voxel während des Trainings zur Optimierung des Modells nutzt.

Als zusätzliche räumliche Features wurden hier die MNI-Koordinaten der Voxel und die Läsionswahrscheinlichkeit (LP, engl. lesion probability) auf Voxelebene verwendet. Diese Features wurden einzeln und in Kombination zu den anderen Features (ADC und PWI-Parameter), die schon beim vorigen Ansatz verwendet wurden, in die Modellierung eingespeist. Insgesamt wurden damit Modelle für die folgenden Featurekombinationen trainiert:

1. DWI- (ADC) und PWI- (CBF, CBV, MTT, T_{max}) Features
2. DWI- und PWI-Features und MNI-Koordinaten
3. DWI- und PWI-Features und LP
4. DWI- und PWI-Features, MNI-Koordinaten und LP

3.5.1 Räumliche Features

Nachdem schon für den lokalen Ansatz alle MRT-Datensätze auf einen MNI-Atlas registriert wurden, sind die MNI-Koordinaten der einzelnen Voxel bereits im aufbereiteten Datensatz vorhanden. Der zweite Anreicherungsschritt der Trainingsdaten für die lokalen Modelle beruht auf der Prämisse von der Symmetrie des Gehirns. Um diese Annahme in Bezug auf den bewusst symmetrisch gewählten MNI-Raum zu nutzen, wurde für diesen Anreicherungsschritt die x-Achse in die Mitte des MNI-Raums verschoben, sodass die hemisphärische Fissur der Ebene für x = 0 entspricht. Damit lie-

ßen sich rechnerisch die Hemisphären durch das Vorzeichen der x-Koordinate unterscheiden. Zur Zusammenlegung der Trainingsdaten von unterschiedlichen Gehirnhälften entfiel dann das Vorzeichen der x-Koordinate für die lokalen Trainingsdaten. Damit wurde bei MNI-Positionen mit denselben Koordinaten nicht mehr zwischen den Hemisphären unterschieden. Beibehalten wurde diese Transformation der x-Koordinate für die MNI-Koordinaten, die in diesem Ansatz als Features verwendet wurden.

Zur Berechnung der LP-Karten wurde jeweils die relative Häufigkeit von Läsionen an allen MNI-Positionen berechnet. Dieses Feature wurde patientenspezifisch für das in dieser Arbeit verwendete LOPO-Kreuzvalidierungsschema erstellt, sodass ein Überlappen von Trainings- und Testdaten vermieden wurde. Dazu wurde für jede Voxelposition die relative Häufigkeit der Läsionen aller anderen Patienten berechnet. Läsionsinformationen aus dem Datensatz, für den die LP-Karte generiert wurde, wurden nicht in diese Berechnung einbezogen, da diese das Ziel der Prädiktion sind und daher bei der Anwendung des Modells unbekannt.

3.5.2 Modelltraining

Als Algorithmen wurden die logistische Regression, Random Forest und XGBoost verwendet. Weil es für eine Modellvalidierung mithilfe von Testdaten zwecks Verzerrung essenziell ist, dass keine Informationen aus den Testdatensätzen in den Prozess des Modelltrainings eingehen, wurde das LP-Feature aus dem globalen Ansatz für jeden Kreuzvalidierungsdurchlauf für alle Patienten individuell berechnet.

Für jedes Random-Forest-Modell wurden 100 Entscheidungsbäume auf jeweils 60 % zufällig gezogener Trainingsdaten trainiert. Während für die meisten Software-Implementierungen des Random-Forest-Algorithmus bekannt ist, dass sie relativ stabile Prognoseergebnisse für die Standardeinstellungen liefern [Probst et al., 2019] und mit zunehmender Menge an Entscheidungsbäumen eher robuster werden, gilt dies nicht für Gradient Boosting Machines bzw. XGBoost. Für diese spielen in Anbetracht des Bias-Varianz-Tradeoffs verschiedene Hyperparameter und insbesondere die Anzahl der Entscheidungsbäume eine Rolle.

Im Sinne der Reproduzierbarkeit und der grundlegenden Hypothese, dass die räumlichen Features in erster Linie die Prognosen positiv beeinflussen, wurde von einer ausführlichen Hyperparameteroptimierung für XGBoost abgesehen. Stattdessen wurde zunächst eine Auswahl plausibler Werte für einzelne Hyperparameter getroffen,

aus denen dann alle möglichen Kombinationen gebildet wurden. Anschließend wurde aus diesem Gitter eine Zufallsauswahl von sieben Hyperparameter-Settings für diese Arbeit gezogen (siehe Tabelle A1 im Anhang), sodass insgesamt 28 XGBoost-Modelle trainiert wurden (sieben Hyperparametersettings pro Featurekombination). Die einzige Konstante zwischen diesen Settings war, dass jeweils nur zehn Bäume trainiert wurden, um gleichzeitig eine kurze Trainingszeit zu gewährleisten und eine Überanpassung an die Trainingsdaten zu vermeiden. Alternativ besteht die Möglichkeit bei XGBoost die Anzahl der Bäume dynamisch in Bezug auf ein Validierungskriterium über ein sogenanntes *early stopping* zu steuern, was jedoch je nach Hyperparametersetting zu deutlich längeren Trainingszeiten führen kann.

Insgesamt wurden damit 36 Modelle zur Validierung dieses Ansatzes trainiert (vier logistische Regressionen, vier Random Forests und 28 XGBoost-Modelle). Da unausgewogene Trainingsdatensätze oft zu schlechteren Ergebnissen führen, wurde für jeden Patienten die Anzahl an Nichtläsionsvoxeln in den Trainingsdaten per Zufallsauswahl auf die Anzahl an Läsionsvoxeln des jeweiligen Patienten reduziert [Jonsdottir et al., 2009; Winder et al., 2019].

Im Gegensatz zum lokalen Ansatz wurde hier für die Integration der räumlichen Features die speedRF-Random-Forest-Implementierung aus dem h2o-R-Paket gewählt [LeDell et al., 2018], weil diese sowohl schneller als auch arbeitsspeichereffizienter ist. Beim lokalen Ansatz wurde sie nicht verwendet, da diese gegenüber der Standard-R-Implementierung eine Konvertierung in ein spezielles Datenformat benötigt. Dies stellt sich aufgrund der zusätzlichen Rechenzeit bei der Vielzahl der Modelle beim lokalen Ansatz als nachteilig heraus. Inhaltlich besteht der größte Unterschied zwischen den beiden Implementierungen darin, dass im randomForest-Paket das Prinzip Ziehen mit Zurücklegen und bei H2O stattdessen Ziehen ohne Zurücklegen bei der zufälligen Datenauswahl für einzelne Bäume angewendet wird. Dies sollte jedoch nur einen minimalen Unterschied in der Praxis ausmachen.

3.6 Evaluierung

Beide Ansätze (lokal und global mit räumlichen Features) wurden im Rahmen einer LOPO-Kreuzvalidierung evaluiert. Dies ist insbesondere für das Training der lokalen Modelle von Vorteil, da die Menge der Trainingsdaten für ein lokales Modell der Anzahl an Patientendatensätzen entspricht. Da der Schwellwert in dieser Arbeit mit der Metrik des Dice-Koeffizienten definiert wird, wurde pro Modell und Kreuzvalidierungsdurch-

lauf jeweils der Schwellwert gewählt, der zum besten Ergebnis des Dice-Koeffizienten auf den Trainingsdaten führt. Hierfür wurde in jedem Durchlauf der Kreuzvalidierung zunächst der optimale Dice-Schwellwert auf den Prognosen der Trainingsdaten (mit bekanntem Outcome) ermittelt. Dies erfolgte durch die Auswertung des Dice-Koeffizienten für alle Schwellwerte von null bis eins in Schritten von 0,01. Anschließend konnte für die Testdaten mit dem resultierenden optimalen Schwellwert zunächst eine Binarisierung der Prognosen stattfinden, um dann basierend auf den Wahrheitswerten der jeweiligen Prognosen (vgl. Confusion Matrix) verschiedene schwellwertbasierte Metriken auszuwerten.

3.6.1 Validierung des lokalen Ansatzes

Der lokale Ansatz wurde im Rahmen einer LOPO-Kreuzvalidierung evaluiert, indem die ROC AUC, der Dice-Koeffizient, die Sensitivität und die Spezifität mit dem globalen und dem hybriden Ansatz verglichen wurden. Der Schwellwert für die Berechnung der Sensitivität und der Spezifität wurde wie beschrieben am Dice-Koeffizienten bestimmt.

3.6.2 Validierung des globalen Ansatzes mit räumlichen Features

Auch der globale Ansatz mit räumlichen Features wurde ebenfalls im Rahmen einer LOPO-Kreuzvalidierung durch eine Auswertung der ROC AUC und den Dice-Koeffizienten evaluiert.

3.7 Feature Importance

In den nächsten Abschnitten wurde überprüft, welchen Einfluss die Features auf die verschiedenen Modellierungen hatten. Dies ermöglicht es u. a., die Relevanz der räumlichen Informationen auf die Prognosegüte der Modelle und deren einzelner Prognosen zu bewerten. Damit lässt sich die Grundhypothese dieser Arbeit, dass räumliche Informationen die Prognosegüte für das Gewebeoutcome verbessern, detailliert auf einzelne (räumliche) Features zurückverfolgen. Ebenso gibt dies Aufschluss über die absolute und relative Wichtigkeit der einzelnen Features und erlaubt die Erklärung der einzelnen Prognosen auf Voxelebene anhand einer Zerlegung in die Anteile der einzelnen Features. Somit lässt sich genau bestimmen, welches Feature wie stark zur Verbesserung eines Modells beiträgt.

3.7.1 Einfluss der Features bei der lokalen logistischen Regression

Da die räumlichen Informationen bei diesem Ansatz implizit über die Position der lokalen Modelle genutzt wurden, wurden zur Überprüfung der Hypothese, wonach die lokalen Modelle räumlich variieren, die Koeffizienten der lokalen logistischen Regression getrennt nach Gehirnregionen[26] ausgewertet und zwischen diesen sowie mit den Koeffizienten der globalen logistischen Regression verglichen. Da einige der Koeffizienten nicht normalverteilt waren, wurden der Median und die mittlere absolute Abweichung vom Median für diese Vergleiche betrachtet.

Für die globalen XGBoost-Modelle, in denen die räumlichen Features als Variablen verwendet wurden, konnten diese mit den anderen Features verglichen werden.

3.7.2 Einfluss der Features in XGBoost-Modellen

Beim globalen Ansatz mit räumlichen Features wurde – stellvertretend für die treebasierten Ensemblemodelle – der Einfluss der Features im jeweils besten XGBoost-Modell (pro Featurekombination) anhand verschiedener Metriken analysiert. Hierzu wurde beim Modelltraining der Gain betrachtet. Welchen Effekt einzelne Features auf die Zielmetriken (ROC AUC und Dice-Koeffizient) der Testdaten haben, wurde anhand der Permutation Importance überprüft, während die Auswirkungen einzelner Voxelfeatures auf die Prädiktion der Testdaten mit den Shapley-Werten untersucht wurden.

Da die Berechnung dieser Metriken zur Feature Importance im Rahmen der LOPO-Kreuzvalidierung erfolgt ist, wurden die für jedes Modell erzielten Ergebnisse für jedes Feature über alle Patienten gemittelt. Anschließend wurden die Features pro Metrik (und Modell) sortiert, sodass eine gewichtete Gesamtreihenfolge der Features erstellt werden konnte. Dadurch weiß man, welche Features am meisten zum Erfolg eines Modells beitragen, und ob sie auch in weiteren Studien eingesetzt werden sollen.

Gain

Der sogenannte Gain beschreibt bei XGBoost die Veränderung der Zielfunktion durch eine zusätzliche Unterteilung (engl. split) eines Pfades eines Entscheidungsbaums beim Modelltraining [Li et al., 2019]. Die Zielfunktion besteht aus zwei Summanden für die Güte der Prädiktion und die Komplexität des Modells. Bei einem negativen Gain

[26] Die betrachteten Gehirnregionen flossen u. a. bei [Winder et al., 2019] als eigenständiges Feature in die Modellierung ein.

erhöht sich die Komplexität stärker als es die Verbesserung des Modells rechtfertigt. Auf Basis dieses Kriteriums kann XGBoost aus verschiedenen Splits auswählen und unnötige Splits vermeiden.

Der so ermittelte Beitrag eines Splits zur Zielfunktion bietet eine einfache und effiziente Methode zur Bewertung des Einflusses einzelner Features auf das Modelltraining. XGBoost summiert den Gain für alle Splits eines Features und gibt diesen für jedes Feature als Feature Importance (in Prozent) aus. Damit kennt man die Wichtigkeit jedes Features zur Optimierung eines Modells.

Shapley-Werte

Shapley-Werte wurden erstmals 1953 von Lloyd Stowell Shapley im Rahmen der kooperativen Spieltheorie eingeführt. Mit diesem Konzept lässt sich eine Belohnung gerecht unter einer Koalition von kooperativen Spielern aufteilen, denn jeder Spieler erhält dadurch genau den Anteil, der seinem Beitrag zum durch die Koalition generierten Überschuss entspricht [Shapley, 1953]. Dieses Konzept wurde kürzlich von [Lundberg und Lee, 2017] auf das Machine Learning übertragen und in mehreren Softwarepaketen wie u. a. [T. Chen et al., 2017] implementiert. Hier entspricht die Rolle der Belohnung (einschließlich des Überschusses der Koalition) einer Vorhersage. Die Spieler korrespondieren mit den Features und der Beitrag eines Spielers steht für die Ausprägung eines Features. Insofern bieten Shapley-Werte ein Konzept, die Vorhersage eines Machine-Learning-Modells in einzelne Featurebeiträge zu zerlegen. Technisch gesehen entspricht die Definition dem durchschnittlichen marginalen Beitrag eines Featurewerts über alle möglichen Featurekombinationen. Shapley-Werte bieten eine ähnliche Interpretationsgrundlage auf Prädiktionsebene wie ein Feature-Koeffizient mit seiner Feature-Ausprägung im Kontext eines linearen Modells. Dies erleichtert die Interpretierbarkeit von komplexen Modellen wie u. a. XGBoost zusätzlich zu ihren Vorteilen gegenüber linearen Modellen, nichtlineare Zusammenhänge und Interaktionen zwischen Features selbstständig zu erlernen.

Um herauszufinden, wie stark die einzelnen Features das Outcome auf den Testdaten differenzieren, wurden für jeden Patienten die durchschnittlichen Shapley-Werte getrennt nach Gewebeoutcome berechnet. Features mit einer besonders großen Differenz in diesen Werten tragen in hohem Maße zur Unterscheidung des Outcomes durch XGBoost bei. Damit spielen solche Features bei der Prädiktion neuer Datensätze eine große Rolle, weshalb sie besonders kritisch überprüft werden müssen.

Permutation Importance

Die Grundidee zur Messung der Permutation Importance für Features geht auf [Breiman, 2001] zurück. Um den Einfluss der Features auf den Dice-Koeffizienten und die ROC AUC zu untersuchen, wurden die Modelle auf Testdaten evaluiert, bei denen zuvor je ein Feature zufällig permutiert wurde. Hierdurch verlor das jeweilige Feature an jeglichem Informationsgehalt. Die drei MNI-Koordinaten wurden in der Auswertung der Permutation Importance als einzige als gemeinsames Feature betrachtet und damit nicht pro Achse (x, y und z), sondern als Koordinatenpunkte permutiert. Damit ist gleichzeitig sichergestellt, dass durch die Permutation das Feature zwar informationslos ist, jedoch keine Koordinaten erzeugt werden, die nicht im Gehirn liegen. Bei der Auswertung wurden für jedes Feature diejenigen Differenzen in den Metriken (ROC AUC, Dice-Koeffizient) berechnet und gemittelt, die für informative und nichtinformative (permutierte) Informationen erzielt wurden.

Rangfolge

Damit sich die Beiträge der Features für die XGBoost-Modelle miteinander vergleichen lassen, wurde das folgende Vorgehen für jedes Modell und jede Feature-Importance-Metrik (Gain, Shapley-Werte und Permutation-Importance bzgl. ROC-AUC- und Dice-Koeffizienten) angewendet: Zunächst wurden die Beiträge der einzelnen MNI-Koordinaten zu einem Gesamt-MNI-Feature aggregiert. Hierzu wurden für den Gain und die Shapley-Werte jeweils die Summen der einzelnen Beiträge der MNI-Koordinaten gebildet. Im Falle des Gain wurde dabei berücksichtigt, dass die Informationen aus den Koordinaten-Achsen tatsächlich orthogonal und komplementär sind. Für Shapley-Werte ist eine Addition von verschiedenen Features zu einem Gesamtfeature aufgrund der Additivität der Shapley-Werte naturgemäß. Für die Permutation Importance wurden die MNI-Koordinaten bereits für die Berechnung der Metrik zu einem Feature zusammengefasst.

Danach wurden für jedes Modell die Features nach ihren Werten bzgl. der einzelnen Metriken sortiert, sodass 16 Rankings (vier Modelle, vier Metriken) erstellt wurden. Im Falle der Shapley-Werte wurden die Features nach der durchschnittlichen Differenz zwischen den Shapley-Werten für Läsions- und für Nichtläsionsvoxel sortiert. Anschließend wurden die Rankings normalisiert und dazu entsprechend der Anzahl der Features pro Modell gleichmäßig zwischen 0 (niedrigster Rang) und 1 (höchster Rang)

skaliert. Abschließend wurde der durchschnittliche Rang für jedes Feature berechnet, sodass ein gewichtetes Gesamtranking über alle Features erstellt werden konnte.

Ob der Effekt von zusätzlichen räumlichen Features und des spezifischen Algorithmus zu einer signifikanten Verbesserung der Modellierung dient, wird in einer weiteren statistischen Analyse getestet.

3.8 Statistische Analyse

3.8.1 Patientenstatistik

Um Auffälligkeiten in den Patientendaten zwischen den verschiedenen Standorten der I-KNOW-Studie zu ermitteln, wurden die Verteilungen der wichtigsten Merkmale auf Unabhängigkeit bzgl. des behandelnden Standorts untersucht. Hierzu wurden der Chi-Quadrat-Test (Geschlecht, Infarktseite, Behandlung mit intravenösem Gewebe-Plasminogen-Aktivator (IV-tPA) – ja/nein), die ANOVA (Alter, NIHSS) und der Kruskall-Wallis-Test (Follow-up-Läsionsvolumen) verwendet (vgl. Abschnitt 4.1).

3.8.2 Modellstatistiken

Für beide Forschungsansätze zur Modellierung des Gewebeoutcomes wurden die Läsionsvorhersageergebnisse der Modelle mithilfe von zweiseitigen, gepaarten T-Tests systematisch auf Unterschiede in den ROC AUC und Dice-Koeffizienten getestet. Für die Sensitivität und Spezifität wurden deren Summen auf Unterschiede geprüft, da diese beiden Metriken gegenläufig sind und daher zusammen betrachtet werden sollten. Alle statistischen Tests verwendeten ein Alpha von 0,05 und wurden mit R (Version 3.4.2) berechnet. Eine Korrektur für multiples Testen wurde nicht durchgeführt.

Statistik für den lokalen Ansatz

Bei den sechs Modellen, die im Rahmen des lokalen Modellierungsansatzes trainiert wurden, diente der Algorithmus als Distinktionsmerkmal (logistische Regression, Random Forest). Getestet wurden die Algorithmen auf Unterschiede bzgl. des Ansatzes (lokal vs. global, lokal vs. hybrid und global vs. hybrid). Zudem fand eine Kategorisierung nach Ansatz (lokal, hybrid, global) statt, wobei dann die Unterschiede bzgl. des verwendeten Algorithmus getestet wurden. Insgesamt wurden pro Metrik neun Tests durchgeführt. Das beste Modell in jeder Metrik (inkl. Sensitivität und Spezifität) wurde

anschließend mittels einseitigem, gepaarten T-Test auf signifikante Unterschiede gegenüber allen anderen Modellen untersucht (vgl. Abschnitt 4.2).

Statistik für den globalen Ansatz mit räumlichen Features

Analog wurden für den globalen Ansatz mit räumlichen Features die zwölf Modelle (drei Algorithmen, vier Featurekombinationen) je einmal nach Algorithmus unterteilt, um Unterschiede bzgl. der Featurekombination zu ermitteln (18 Vergleiche, sechs pro Algorithmus). Danach wurden die Modelle nach Featurekombination gruppiert, um Unterschiede bzgl. des verwendeten Algorithmus festzustellen (12 Vergleiche, drei pro Featurekombination). Insgesamt wurden 60 Tests durchgeführt (30 pro Metrik). Das beste Modell in jeder Metrik wurde zusätzlich mittels einseitigem gepaarten T-Test gegenüber allen anderen Modellen auf signifikante Unterschiede untersucht (vgl. Abschnitt 4.3).

Statistik für den Vergleich zwischen dem lokalen Ansatz und globalen Ansatz mit räumlichen Features

Auch zwischen den beiden Forschungsansätzen wurden die Modelle mittels zweiseitigen, gepaarten T-Tests verglichen, um zu prüfen, welche Art der Integration räumlicher Informationen vorteilhafter ist. Da der erste Forschungsansatz nur für die logistische Regression und Random Forest evaluiert wurde, war die Berücksichtigung von XGBoost-Modellen nicht erforderlich. Hierbei wurde jedes der beiden lokalen und hybriden Modelle mit den je drei globalen Modellen mit räumlichen Informationen des korrespondierenden Algorithmus verglichen, sodass zwölf Modellvergleiche bzw. insgesamt 24 Tests (zwei pro Metrik) durchgeführt wurden (vgl. Abschnitt 4.4).

Vier weitere Tests (zwei pro Metrik) wurden durchgeführt, um die unterschiedlichen Samplingverfahren (stratifiziertes nach Outcome vs. Downsampling der Läsionsvoxel pro Patienten) gegenüberzustellen, wozu die jeweiligen globalen Modelle, die keine räumlichen Informationen beinhalteten, miteinander verglichen wurden. Diese Gegenüberstellung soll im Wesentlichen darüber Aufschluss geben, welches der beiden unterschiedlichen Samplingverfahren, die in dieser Arbeit genutzt wurden, sich besser zur Auswahl der Trainingsdaten der Modelle eignet (vgl. Abschnitt 4.4).

4 Ergebnisse

4.1 Stichprobe

Insgesamt erfüllten 99 von 180 vorhandenen Datensätzen aus der I-KNOW-Studie die Einschlusskriterien dieser Arbeit. Von den 180 Patienten mussten 29 wegen fehlender klinischer Daten und acht Patienten wegen fehlender Bilddaten ausgeschlossen werden. Die verbleibenden 143 Datensätze wurden nochmal um 30 Datensätze aus folgenden Gründen bereinigt: fehlende FU-FLAIR (sechs Datensätze), unbekannte Infarktzeit (vier), fehlende Bilddaten zum Aufnahmezeitpunkt (vier), verzerrte PWI-Sequenzen (zwei), kein Kontrastmittel in den PWI-Sequenzen (zwei), kein akutes ischämisches Defizit (zwei), bilaterales ischämisches Defizit (zwei), verzerrte DWI-Sequenzen (zwei), fehlende PWI-Sequenzen (einer), Läsion in der Nähe eines alten Infarkts (einer), verzerrte FU-FLAIR (einer), Rotationsartefakte in der DWI-Sequenz (einer), fehlende Schichten in der DWI-Sequenz (einer), fehlerhaft angegebene Zeiten für die PWI-Sequenzen (einer). Zusätzlich wurden 14 Datensätze nicht ins Sample aufgenommen, weil sie nicht einwandfrei mittels AnToNIa prozessiert werden konnten.

Die zentralen Merkmale der 99 in dieser Arbeit berücksichtigten Patienten sind in Tabelle 2 dargestellt. Die Patienten ähneln sich über die verschiedenen Standorte hinweg in Bezug auf die Geschlechterverteilung ($p = 0,92$), die Verteilung der betroffenen Infarktseiten ($p = 0,50$), das durchschnittliche Alter ($p = 0,33$) und den Median NIHSS ($p = 0,25$). Im Gegensatz dazu fallen die Follow-up-Läsionsvolumina an den beiden Standorten mit den wenigsten Patienten (Cambridge und Aarhus) deutlich geringer aus als an den anderen Standorten ($p = 0,14$). An beiden Standorten wurden alle Patienten mit IV-tPA behandelt ($p < 0,01$).

Tabelle 2: Eigenschaften der 99 Patienten aus der I-KNOW-Studie

Standort	N	Geschlecht (m/w)	Ø Alter	Infarktseite (links/rechts)	IV-tPA (ja/nein)	Ø NIHSS	Median Follow-up-Läsionsvolumen (in ml)
Aarhus	12	7/5	65,2±7,9	6/6	12/0	9,1±5,3	16,49±30,67
Cambridge	2	1/1	74,5±5	0/2	2/0	8,5±5	7,17±1,02
Girona	28	14/14	68,7±12,1	15/13	14/14	12,6±6,1	36,99±87,25
Hamburg	19	10/9	70,0±12,5	11/8	16/3	9,1±4,1	27,98±49,23
Lyon	38	23/15	64,0±14,1	24/14	19/19	11,2±6,6	31,62±93,48
Gesamt	99	55/44	66,8±12,6	56/43	63/36	10,9±6	28,62±79,26
Signifikanz		p = 0,920	p = 0,330	p = 0,495	p = 0,001	p = 0,252	p = 0,139

Unterschiede in Bezug auf den behandelnden Standort wurden mittels Chi-Quadrat-Test (Geschlecht, Infarktseite, IV-tPA), ANOVA (Alter, NIHSS) und Kruskall-Wallis-Test (Follow-up-Läsionsvolumen) getestet (Quelle: in Anlehnung an Grosser et al., 2020b, ergänzt um p-Werte).

4.2 Lokaler Ansatz

Die Voraussetzungen von mindestens drei Läsions- und drei Nichtläsionsvoxeln für das Training eines lokalen Modells erfüllten insgesamt 72,33 % der als Gewebe segmentierten MNI-Positionen. Für die Anzahl an lokalen Modellen spielten die Anreicherungsschritte der lokalen Trainingsdaten eine entscheidende Rolle, weil ansonsten nicht genügend MNI-Positionen als Datengrundlage vorhanden gewesen wären. Ohne die Anreicherung hätten lediglich an 32,27 % der MNI-Positionen lokale Modelle trainiert werden können. Wie die Anreicherung erfolgt ist, wird im nächsten Abschnitt erklärt.

4.2.1 Anreicherung der lokalen Trainingsdaten

Durch das Anreichern von bis zu zwei Datenpunkten für jeden Patienten innerhalb eines Suchradius von $\sqrt{5}$ mm um die Modellpositionen konnten bereits an 59,58 % der MNI-Positionen lokale Modelle trainiert werden. Durch das zusätzliche Zusammenlegen der Trainingsdatensätze aus unterschiedlichen Gehirnhälften (ohne die Anwendung des Suchradius) lag die Quote bei 46,65 % der MNI-Positionen, an denen lokale Modelle trainiert werden konnten. Standen beim lokalen Ansatz dennoch nicht genügend Trainingsdaten für ein lokales Modell zur Verfügung, dann wurden Vorhersagen für Voxel an diesen MNI-Positionen durch das globale Modell substituiert.

4.2.2 Ergebnisse der Läsionsvorhersage

Die quantitativen Ergebnisse für die drei Modellierungsansätze (global, lokal und hybrid) und die zwei erprobten Machine-Learning-Algorithmen (logistische Regression und Random-Forest) sind in Tabelle 3 anhand der durchschnittlichen ROC AUCs, Dice-Koeffizienten sowie Sensitivitäts- und Spezifitäts-Metriken aufgeschlüsselt. Die Ergebnisse der statistischen Tests auf signifikante Unterschiede zwischen den Ansätzen und Algorithmen sind in Tabelle A2 im Anhang aufgeführt.

Für beide Algorithmen führt der hybride Ansatz zu mindestens signifikanten Verbesserungen ($p < 0,05$) gegenüber dem lokalen und dem globalen Ansatz, was sich am Dice-Koeffizienten und der ROC AUC zeigt. Gleiches gilt – bis auf den Vergleich mit der lokalen logistischen Regression – für die Summe aus Sensitivität und Spezifität.

Tabelle 3: Ergebnisse der Modelle beim lokalen und hybriden Ansatz

Modell	Lokalität	Ø ROC AUC	Ø Dice-Koeffizient	Ø Sensitivität	Ø Spezifität	Ø Vorhersagezeit (s)
LR	Global	0,809±0,110**	0,322±0,218**	0,386±0,243**	0,959±0,047	**0,055±0,031**
LR	Local	0,861±0,109**	0,337±0,221**	**0,448±0,254**	0,955±0,041**	0,062±0,015*
LR	Hybrid	**0,872±0,092**	0,348±0,221	0,444±0,252	0,955±0,047**	0,126±0,038**
RF	Global	0,789±0,104**	0,319±0,215**	0,361±0,218**	**0,965±0,037**	29,762±6,857**
RF	Local	0,845±0,099**	0,311±0,208**	0,404±0,208**	0,956±0,030**	704,859±146,593**
RF	Hybrid	0,859±0,089**	**0,353±0,220**	0,415±0,231**	0,964±0,034	736,284±148,309**

Durchschnittliche Werte für die ROC AUCs, die Dice-Koeffizienten, die Sensitivität, die Spezifität und die Vorhersagezeiten im Rahmen der Leave-One-Patient-Out-Kreuzvalidierung für jedes der Modelle. Die besten Ergebnisse in Bezug auf die einzelnen Metriken sind jeweils fett gedruckt hervorgehoben. Signifikante Unterschiede einzelner Modelle zum jeweils besten Modell bzgl. der jeweiligen Metrik wurden mit einem einseitigen T-Test berechnet und mit einem Stern (*) bei einem Signifikanzniveau von $p < 0,05$ bzw. mit zwei Sternen (**) bei einem Signifikanzniveau von $p < 0,01$ gekennzeichnet. Nominale p-Werte wurden ohne Korrektur für multiples Testen berechnet, ähnlich wie in [Maier et al., 2015]. LR = logistische Regression, RF = Random Forest (Quelle: Grosser et al., 2020b).

Die Vorhersagen des lokalen Ansatzes zeigten hochsignifikante ($p < 0,01$) Verbesserungen hinsichtlich der ROC AUC gegenüber den Vorhersagen des globalen Modells. Zusätzlich ergaben sich mindestens signifikante Verbesserungen in der Summe aus Sensitivität und Spezifität, die auf den deutlich erhöhten Erkennungsraten von Läsionsvoxeln (+0,06 für die logistische Regression und +0,04 für Random Forest) beruhen. Beim Dice-Koeffizienten trat keine signifikante Verbesserung für die logistische Regression ($p = 0,07$) ein, aber eine kleine, jedoch nicht signifikante Verschlechterung für den Random Forest.

Die höchsten Durchschnittswerte für die ROC AUC bzw. den Dice-Koeffizienten erreichten die hybride logistische Regression (0,872±0,092) und der hybride Random Forest (0,353±0,220). Die Verbesserungen dieser Modelle hinsichtlich der jeweiligen Metrik waren hochsignifikant ($p < 0,01$, einseitiger, gepaarter T-Test) gegenüber allen anderen globalen und lokalen Modellen. Dabei erzielte die hybride logistische Regression Effektstärken von 0,011 und 0,061 im Vergleich zum jeweils besten lokalen und globalen Modell. Das hybride Random-Forest-Modell erreichte einen um 0,016 und 0,025 höheren Dice-Koeffizienten als das jeweils beste lokale bzw. globale Modell.

Der höchste Durchschnitt von ROC AUCs, Dice-Koeffizienten, Sensitivitäts- und Spezifitätswerten wurde für die hybride logistische Regression beobachtet (0,872±0,092; 0,348±0,221; 0,444±0,252; 0,955±0,047). Darauf folgten die lokale logistische Regression, der hybride und der lokale Random Forest sowie die globale logistische Regression und der globale Random Forest.

Während die logistische Regression generell schnellere Vorhersagen als der Random Forest lieferte, war der globale Ansatz stets schneller als der lokale und der hybride Ansatz.

Exemplarisch am Fallbeispiel eines ausgewählten Patienten findet sich in Abbildung 17 eine visuelle Gegenüberstellung der finalen Läsionsvorhersagen der globalen, lokalen und hybriden logistischen-Regressions- und Random-Forest-Modelle. Ersichtlich ist, dass die Vorhersagen der globalen Modelle den schärfsten Kontrast zwischen Läsions- und Nichtläsionsvoxeln erzeugen. Die Vorhersagen des lokalen Ansatzes besitzen dagegen ein etwas stärkeres Rauschen, aber auch eine höhere Sensitivität für Läsionsvoxel. Der hybride Ansatz, der den globalen (hohe Spezifität) und lokalen (hohe Sensitivität) vereint, führt zur insgesamt besten Vorhersage der Infarktläsion.

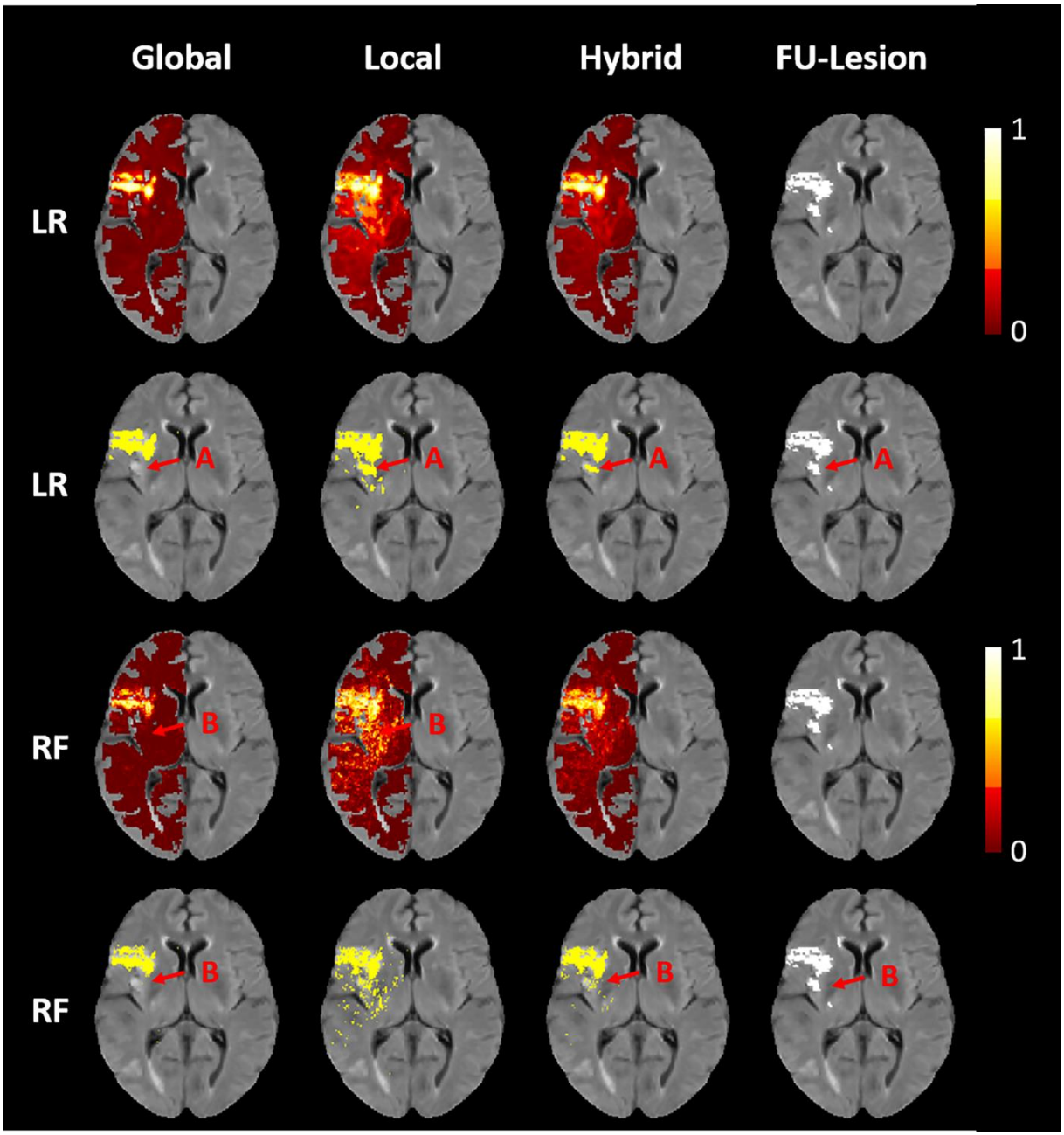

Abbildung 17: Läsionsvorhersage beim lokalen Ansatz für einen ausgewählten Patienten: Abbildungen der finalen Infarktvorhersagen (erste und dritte Reihe) und der binären Läsionsmasken (zweite und vierte Reihe) für die globalen, lokalen und hybriden (von links nach rechts) LR- und RF-Modelle (von oben nach unten) für einen ausgewählten Patienten sowie das dazugehörige tatsächliche Follow-up-Läsionsergebnis (ganz rechts). In dem Bereich, in dem das globale Modell hier die tatsächlichen Läsionen unterschätzt, berechnen die lokalen Modelle höhere Infarktwahrscheinlichkeiten, wodurch die binären Vorhersagen des lokalen Ansatzes mehr mit den tatsächlichen Läsionen übereinstimmen; sowohl für LR (A) als auch für RF (B). Allerdings sind die Läsionsvorhersagen der globalen LR- und RF-Modelle sehr glatt und zusammenhängend, wohingegen der lokale Ansatz, vor allem für RF, zur starken Streuung des angezeigten Infarktrisikos neigt. Dennoch konzentriert sich die Streuung des lokalen Ansatzes auf die tatsächlichen Infarktregionen, sodass im Endergebnis die hybride Vorhersage glatter ist als die lokalen Ansätze und auch zu den insgesamt besten qualitativen Ergebnissen führt (Quelle: Grosser et al., 2020b).

4.2.3 Einfluss der Features bei der lokalen logistischen Regression

Da einige der Koeffizienten der lokalen logistischen Regressionsmodelle keiner Normalverteilung folgten, wurden hierfür die Medianwerte berechnet, die in Tabelle A3 im Anhang dargestellt sind. Daran lässt sich ablesen, dass ein erheblicher Unterschied zwischen den Koeffizienten der lokalen logistischen Regressionsmodelle zwischen den Hirnregionen besteht, was nicht nur den Nutzen der lokalen Modelle unterstreicht, sondern auch auf den unterschiedlichen Zusammenhang zwischen den ADC- und PWI-Parametern mit dem Läsionsrisiko je nach Gehirnregion hinweist.

4.3 Räumliche Features in globalen Modellen

Zum späteren Abgleich der verschiedenen Prognosemodelle sind die diesbezüglichen Werte von ROC AUCs und den Dice-Koeffizienten in Tabelle 4 dargestellt. (Eine vollständige Auflistung dieser Metriken inkl. aller XGBoost-Settings ist in Tabelle A4 angegeben.) Die signifikanten Unterschiede zwischen den Featurekombinationen und Algorithmen, die sich aus den statistischen Tests ergeben haben, sind in Tabelle A5 aufgeführt. Da die Modelle mit dem höchsten Durchschnitt aus ROC AUCs und Dice-Koeffizienten pro Featurekombination zu den zuverlässigsten Voxelvorhersagen führten, wurden diese für XGBoost betrachtet.

4.3.1 Ergebnisse der Läsionsvorhersage

Unabhängig von der Featurekombination wiesen die beiden tree-basierten Ensemblemodelle stets hochsignifikante Verbesserungen für beide Metriken ($p < 0,01$) gegenüber der logistischen Regression bei der Vorhersage auf den Testdaten auf.

XGBoost führte gegenüber Random Forest zu hochsignifikanten Verbesserungen bzgl. der ROC AUC für die Featurekombinationen ADC + PWI und ADC + PWI + LP. Für alle Featurekombinationen bewirkte XGBoost eine Verbesserung im Dice-Koeffizienten, wobei diese für beide Kombinationen mit dem LP-Feature hochsignifikant waren.

Die besten Ergebnisse hinsichtlich der ROC AUC und des Dice-Koeffizienten lieferten zwei XGBoost-Modelle mit räumlichen Features. Dabei wurden eine durchschnittliche ROC AUC von 0,89±0,09 (ADC + PWI + MNI) und ein durchschnittlicher Dice-Koeffizient von 0,40±0,23 (ADC + PWI + LP) erreicht. Das XGBoost-Modell mit den

Features ADC + PWI + LP erreichte den höchsten Durchschnitt aus ROC AUCs und Dice-Koeffizienten.

Beide Modelle zeigten gegenüber Modellen ohne Berücksichtigung von räumlichen Informationen hochsignifikante Verbesserungen ($p < 0,01$, einseitiger T-Test) in den genannten Metriken (siehe Tabelle 4). Die minimalen Effektstärken (gegenüber den besten Modellen ohne räumliche Informationen) betrugen 0,06 (ROC AUC) und 0,05 (Dice-Koeffizient).

Tabelle 4: Ergebnisse der Modelle beim globalen Ansatz mit räumlichen Informationen

Modell	Features	Setting	Ø ROC AUC (Training)	Ø ROC AUC	Ø Dice-Koeffizient	Trainingszeit (s)
LR	ADC + PWI		0,817±0,001	0,813±0,107**	0,317±0,220**	41
LR	ADC + PWI + MNI		0,849±0,001	0,827±0,100**	0,292±0,229**	48
LR	ADC + PWI + LP		0,901±0,001	0,874±0,108**	0,319±0,238**	45
LR	ADC + PWI + MNI + LP		0,902±0,001	0,877±0,099**	0,322±0,232**	44
RF	ADC + PWI		0,887±0,001	0,826±0,104**	0,341±0,218**	614
RF	ADC + PWI + MNI		0,969±0,003	0,891±0,092	0,383±0,226**	628
RF	ADC + PWI + LP		0,950±0,001	0,883±0,104**	0,371±0,227**	622
RF	ADC + PWI + MNI + LP		0,980±0,000	0,889±0,092	0,368±0,228**	758
XGB	ADC + PWI	7	0,845±0,001	0,830±0,105**	0,346±0,220**	77
XGB	ADC + PWI + MNI	3	0,926±0,004	**0,893±0,085**	0,387±0,213	179
XGB	ADC + PWI + LP	7	0,918±0,001	0,888±0,101	**0,395±0,229**	95
XGB	ADC + PWI + MNI + LP	6	0,937±0,001	0,887±0,098*	0,386±0,224	157

Die besten Ergebnisse in Bezug auf die einzelnen Metriken sind jeweils fett gedruckt. Signifikante Unterschiede einzelner Modelle zum jeweils besten Modell bzgl. der jeweiligen Metrik wurden mit einem einseitigen T-Test berechnet und bei einem Signifikanzniveau von $p < 0,05$ mit einem Stern (*) bzw. mit zwei Sternen (**) bei einem Signifikanzniveau von $p < 0,01$ gekennzeichnet. Nominale p-Werte wurden ohne Korrektur für multiples Testen berechnet [vgl. Maier et al., 2015]. Für die XGBoost-Modelle wurde hier jeweils das beste Setting (gemessen am Durchschnitt aus ROC AUCs und Dice-Koeffizienten) pro Featurekombination angegeben. Die Ergebnisse für alle Modelle sind in Tabelle A4 im Anhang zu finden. Die angegebenen Trainingszeiten beziehen sich auf die Zeit zum Training eines einzelnen Modells auf dem gesamten Trainingsdatensatz. LR = logistische Regression, RF = Random Forest, XGB = XGBoost, ADC = apparent diffusion coefficient, PWI = perfusion-weighted imaging parameters, MNI = MNI coordinates, LP = lesion probability (Quelle: in Anlehnung an Grosser et al., 2020a, ergänzt um die ROC AUC auf den Trainingsdaten).

Selbst bei Modellen, die mit demselben Algorithmus trainiert wurden, verbesserten sich alle XGBoost- und Random-Forest-Modelle ($p < 0,01$) durch den Einbezug der räumlichen Features hoch signifikant (siehe Abbildung 18 und Tabelle A5). Allerdings bewirkte die gleichzeitige Hereinnahme von beiden räumlichen Features lediglich für

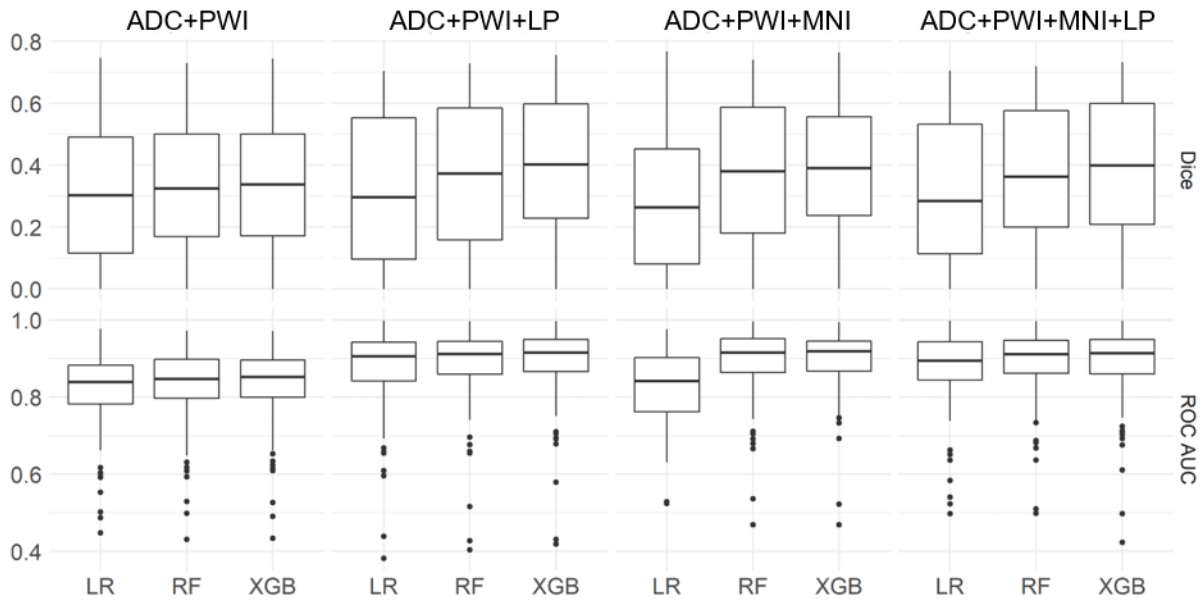

Abbildung 18: Verteilung der Zielmetriken beim globalen Ansatz mit räumlichen Features: Die Boxplots in dieser Abbildung beschreiben die Verteilungen der ROC AUCs und der Dice-Koeffizienten für alle Featurekombinationen und alle LR-, RF- sowie die besten XGB-Modelle (Quelle: Grosser et al., 2020a).

die ROC AUC des Random-Forest-Modells mit den Features ADC + PWI + LP eine signifikante Verbesserung gegenüber einem Modell mit nur einem räumlichen Feature.

Für die logistische Regression führte die Integration des LP-Features stets zu einer hoch signifikanten Verbesserung in der ROC AUC. Bei der Integration der MNI-Koordinaten verschlechterte sich hingegen der Dice-Koeffizient signifikant.

Während die Integration von räumlichen Informationen insbesondere für die tree-basierten Modelle zu Verbesserungen führte, neigten diese allerdings auch verstärkt dazu, sich an die Trainingsdaten anzupassen. Dies lässt sich an der immer größer werdenden Differenz aus der ROC AUC für Vorhersagen auf den Trainings- und Testdaten bei der Hereinnahme weiterer Features ablesen. Während für die logistische Regression mit ADC + PWI diese Differenz lediglich 0,004 betrug und sich um 0,018 (MNI), 0,025 (LP) und 0,021 (MNI + LP) steigerte, war diese Überanpassung für XGBoost schon bei der Verwendung von ADC + PWI etwas stärker (0,015) und stieg bei der Integration der räumlichen Features um 0,018 (MNI) bzw. 0,013 (LP) und um 0,035 (MNI + LP). Besonders stark war die Überanpassung initial beim Random Forest, der auf wesentlich mehr Trees als XGBoost trainiert wurde. So war der Wert für die ROC AUC dieses Modells auf den Trainingsdaten schon für ADC + PWI um 0,061 höher als auf den Trainingsdaten. Bei der Integration der räumlichen Features erhöhte sich dieser Unterschied um 0,017 (MNI), 0,06 (LP) und 0,030 (MNI + LP).

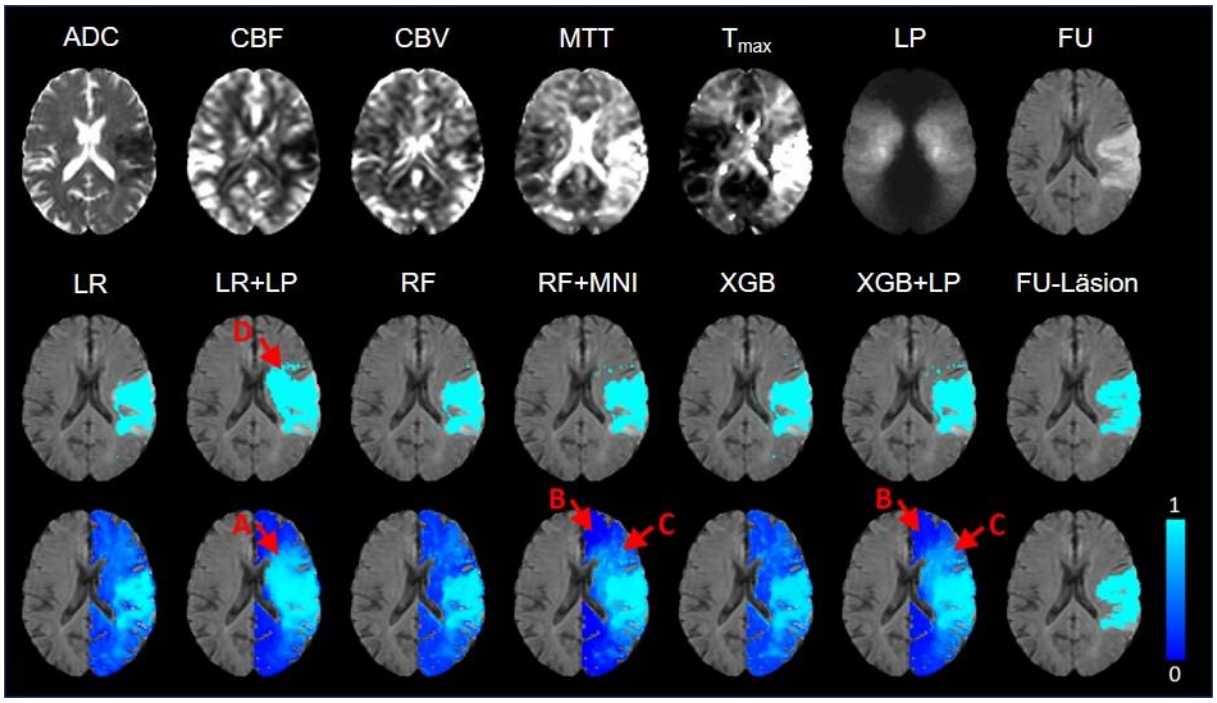

Abbildung 19: Vorhersagebeispiel beim globalen Ansatz mit räumlichen Features: Obere Reihe: Bildgebungsparameter, Läsionswahrscheinlichkeitskarte (LP) und Follow-up-FLAIR-Datensätze. Mittlere und untere Reihe: Vorhersagen in Form von binären Infarktvorhersagen und Prognosescores für Modelle ohne und die besten Modelle mit räumlichen Features sowie die finale Follow-up-Läsion für einen ausgewählten Patienten. Für diesen Patienten stimmen die binären Vorhersagen gut mit dem finalen Infarktergebnis überein, außer für das LR-Modell, das die Läsionswahrscheinlichkeiten verwendet (A). Dabei unterscheiden sich die Vorhersagekarten der Modelle vor allem in der räumlichen Verteilung von Voxeln mit niedrigem (B) und mittlerem (C) Läsionsrisiko. Während diese Risikobereiche bei Modellen, die nur ADC- und PWI-Parameter enthalten, eher zufällig verteilt erscheinen, konzentrieren sich bei Modellen mit räumlichen Informationen die Voxel mit niedrigem Risiko (B) in Bereichen mit geringer Läsionswahrscheinlichkeit und Voxel mit mittlerem Risiko (C) in Bereichen mit hoher Läsionswahrscheinlichkeit. Dies führt zu glatten Vorhersagekarten, sodass die infarzierten Bereiche im Falle von RF und XGB gut unterscheidbar bleiben. Beim LR-Modell zeigt sich allerdings eine starke Überschätzung des Infarkts in der binären Vorhersagekarte (D) (Quelle: Grosser et al., 2020a).

Eine Gegenüberstellung der finalen Läsionsvorhersagen von Modellen ohne räumliche Informationen und den besten Modellen mit räumlichen Features wird am Beispiel eines ausgewählten Patienten in Abbildung 19 veranschaulicht. In dieser ist zu sehen, dass die Vorhersagen mittels der Random-Forest- und XGBoost-Modelle und die Hereinnahme der räumlichen Features für diese Modelle einen durchweg positiven Effekt haben. Lediglich die logistische Regression profitiert nicht durchgängig von der Hereinnahme der räumlichen Features. Während die Verwendung des LP-Features (mit oder ohne MNI-Koordinaten) zwar zu einem starken Anstieg der ROC AUC bei

nur leicht steigendem Dice-Koeffizienten führt, überschätzt die Vorhersage der logistischen Regression den Infarkt im Bereich mit erhöhtem LP-Feature.

4.3.2 Einfluss der Features in XGBoost-Modellen

Gain

Der prozentuale Gain für die Features der besten XGBoost-Modelle (pro Featurekombination) wird in Abbildung 20 dargestellt. Durch die Prozentangaben variieren die Werte vor allem mit der Anzahl der Modellfeatures. Besonders betrifft dies die Werte für die beiden PWI-Features T_{max} und MTT, die Standardabweichungen von 12 % (T_{max}) bzw. 16,1 % (MTT) aufweisen. Der in allen Modellen verwendete ADC trägt beim Modelltraining im Durchschnitt 14,4 % (±5,5 %) zur Verringerung der Zielfunktion bei. Den höchsten Gain erzielten von den ebenfalls stets genutzten PWI-Features T_{max} und MTT für alle Modelle. Deutlich höher als der Gain von MTT lag dabei der Gain von T_{max} in den drei Modellen mit räumlichen Features (insbesondere in den beiden mit LP-Feature). Während sich der Einfluss von CBF mit der Integration von räumlichen Features deutlich verringerte, blieb der Gain für CBV durchgängig niedrig. Für die Featurekombination DWI, PWI und MNI-Koordinaten tragen die einzelnen Koordinaten nahezu identisch knapp über 12 % zum gesamten Gain bei, sodass sich der Anteil der MNI-Koordinaten zu etwa 37 % aufsummiert. Auch das LP-Feature trägt mit 36,2 % den größten Anteil am Gain bei, wenn es als einziges räumliches Feature verwendet wird.

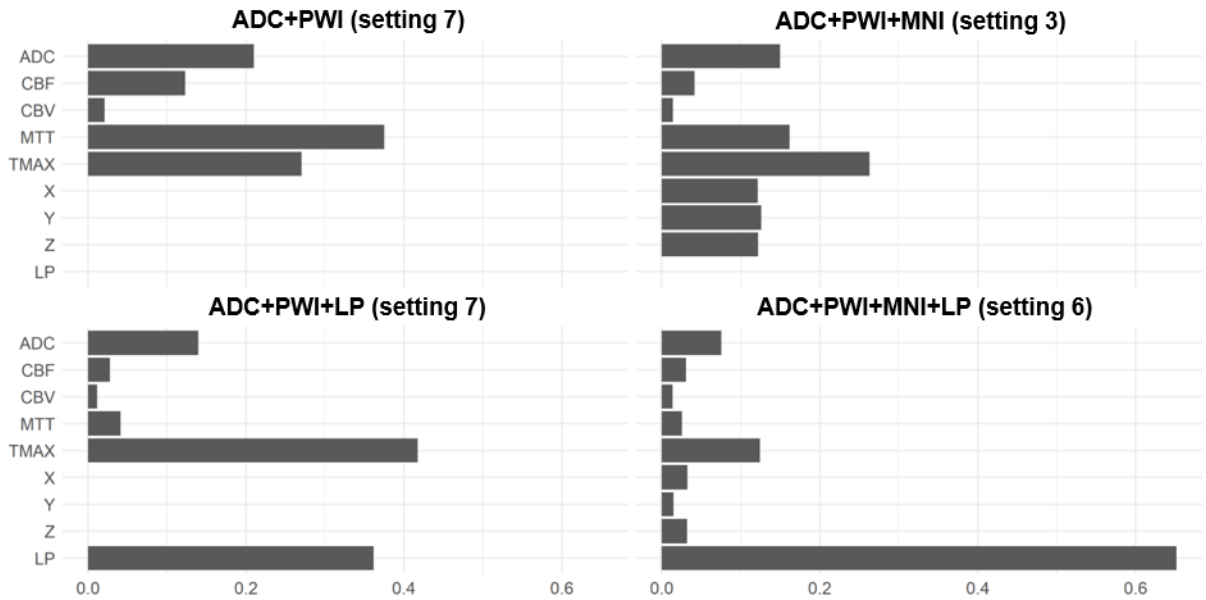

Abbildung 20: Gain pro Feature für die besten XGB-Modelle pro Featurekombination (Quelle: Grosser et al., 2020a).

Bei der gemeinsamen Hereinnahme der MNI-Koordinaten und des LP-Features weist Letzteres mit 65,2 % eindeutig den höchsten Anteil am Gain auf. In diesem Modell generiert XGBoost verhältnismäßig wesentlich weniger Informationen aus den MNI-Koordinaten, sodass diese nur 8 % (x = 3,3 %, y = 3,2 % und z = 1,5 %) des Gains betragen.

Shapley-Werte

Mit dem Gain wurde für jedes Feature angegeben, welchen Beitrag dieses an der Minimierung der Zielfunktionen der Modelle während des Trainings geleistet hat. Um den Einfluss der Features und ihrer Ausprägungen auf Vorhersagen der Modelle zu bestimmen, wurden Shapley-Werte auf Voxelebene für die besten XGBoost-Modelle berechnet (pro Featurekombination; auf den Testdaten im Rahmen der Kreuzvalidierung) (siehe Abbildung 21).

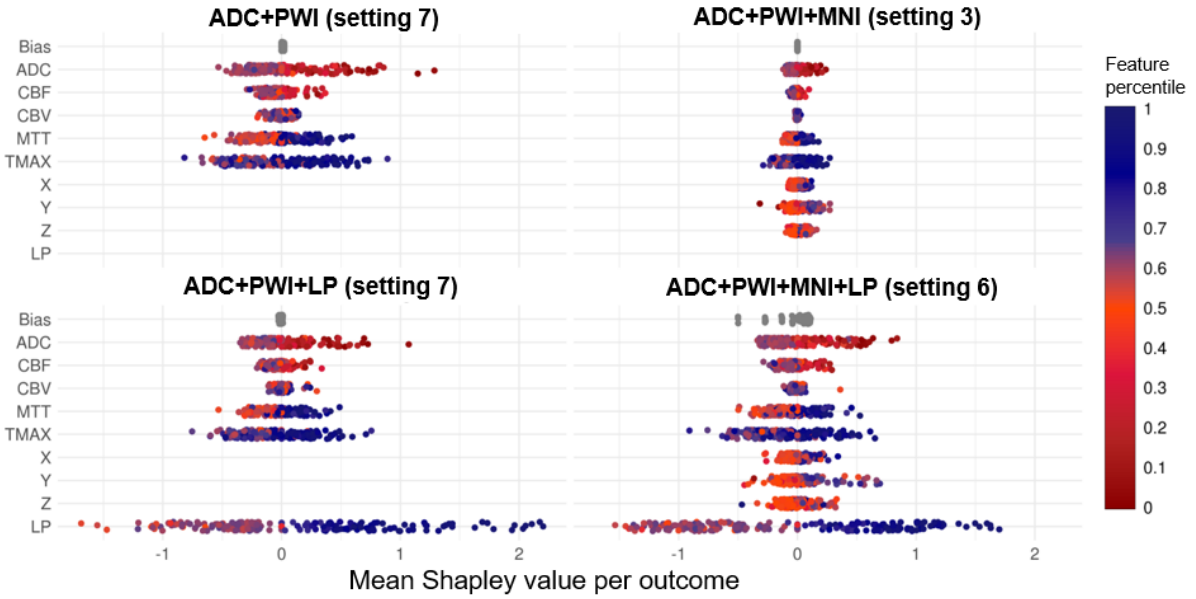

Abbildung 21: Durchschnittliche Shapley-Werte und Featureausprägungen pro Voxel-outcome: Durchschnittliche Shapley-Werte (x-Achse) und Featureausprägungen (in Perzentilen; Farbskala) für die besten XGB-Modelle (pro Featurekombination). Die Mittelwerte wurden einmal pro Outcome (Läsion ja/nein) berechnet. Da der Biasterm in jedem Modell konstant ist, unterscheidet er sich nicht für unterschiedliche Voxeloutcomes und variiert nur aufgrund der einzelnen Kreuzvalidierungsdurchläufe. Die Farbkodierung der Punkte repräsentiert die Mittelwerte der entsprechenden Featureausprägungen (in Läsions- oder Nichtläsionsvoxeln) in Bezug auf ihr Perzentil im Datensatz. So markieren die blauen Punkte in den positiven Bereichen der Shapley-Werte für das LP-Feature, dass die entsprechenden LP-Merkmalswerte im Durchschnitt relativ hoch in Bezug auf die LP-Perzentile in den zugrunde liegenden Datensätzen waren. (Positive Shapley-Werte weisen auf eine positive Korrelation zwischen dem jeweiligen Feature und einer hohen Läsionsrisikovorhersage hin. Der Einfluss auf die endgültige Vorhersage ist jedoch nicht linear, da XGB die Shapley-Werte auf der Logit-Skala berechnet) (Quelle: Grosser et al., 2020a).

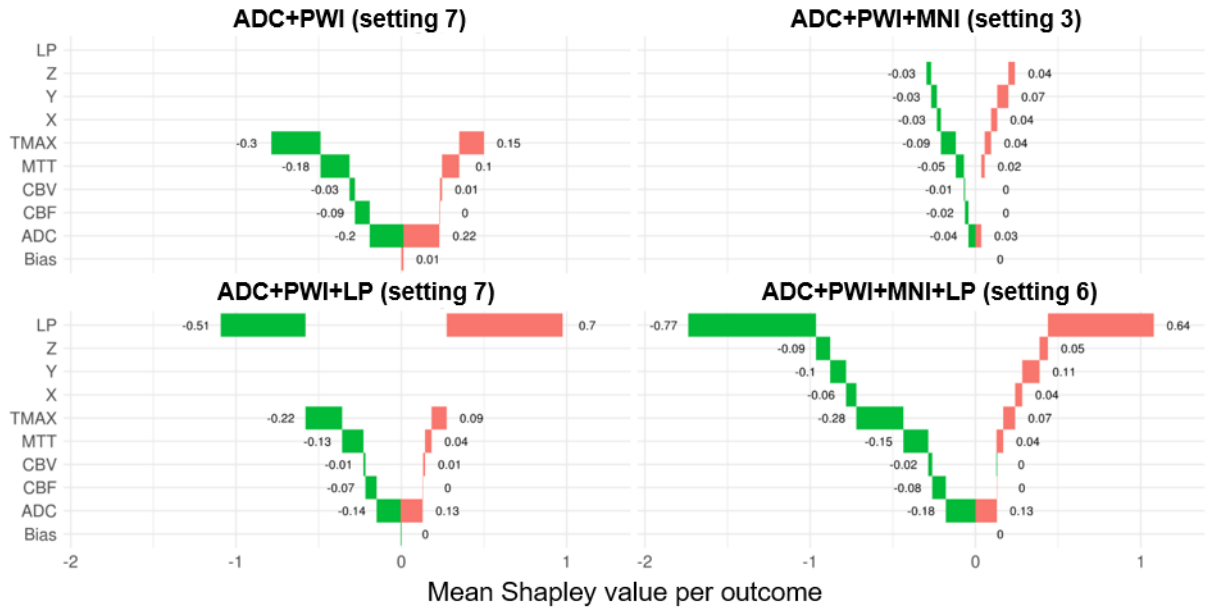

Abbildung 22: Wasserfallaufschlüsselung durchschnittlicher Shapley-Werte pro Voxeloutcome und Modell: Die grünen Balken stehen für die durchschnittlichen Shapley-Werte der Vorhersagen für Nichtläsionsvoxel. Somit entspricht die Summe über alle grünen Balken dem durchschnittlichen Vorhersagewert für Voxel ohne Läsion (auf der Logit-Skala). Dementsprechend stellen die roten Balken die durchschnittlichen Shapley-Werte der Vorhersagen für Läsionsvoxel dar (Quelle: Grosser et al., 2020a).

Shapley-Werte gelten im aktuellen Wissenschaftsdiskurs aufgrund ihrer vorteilhaften Eigenschaften als das beste Interpretationsmittel von Black-Box-Machine-Learning-Modellen. Dazu gehört beispielsweise, dass die Summe der Shapley-Werte aller Features für ein Voxel (einschließlich des Bias) der Läsionsvorhersage des Modells (auf der Logit-Skala) entspricht. Dies ermöglicht es z. B., die Beiträge der einzelnen Features an den Vorhersagen für die unterschiedlichen Outcomeklassen zu beziffern. Für jedes Feature sind in Abbildung 22 demnach die durchschnittlichen Beiträge an den Vorhersagen für Läsions- und Nichtläsionsvoxel visuell aufbereitet.

Da Shapley-Werte für mehrere Features (auf der Logit-Skala) addiert werden können, lassen sich Vorhersagen auf Voxelebene auch in die Beiträge einzelner Features zerlegen (siehe Abbildung 23 für die beste Featurekombination und Abbildung 24 für eine analoge Darstellungsweise der anderen drei untersuchten XGBoost-Modelle).

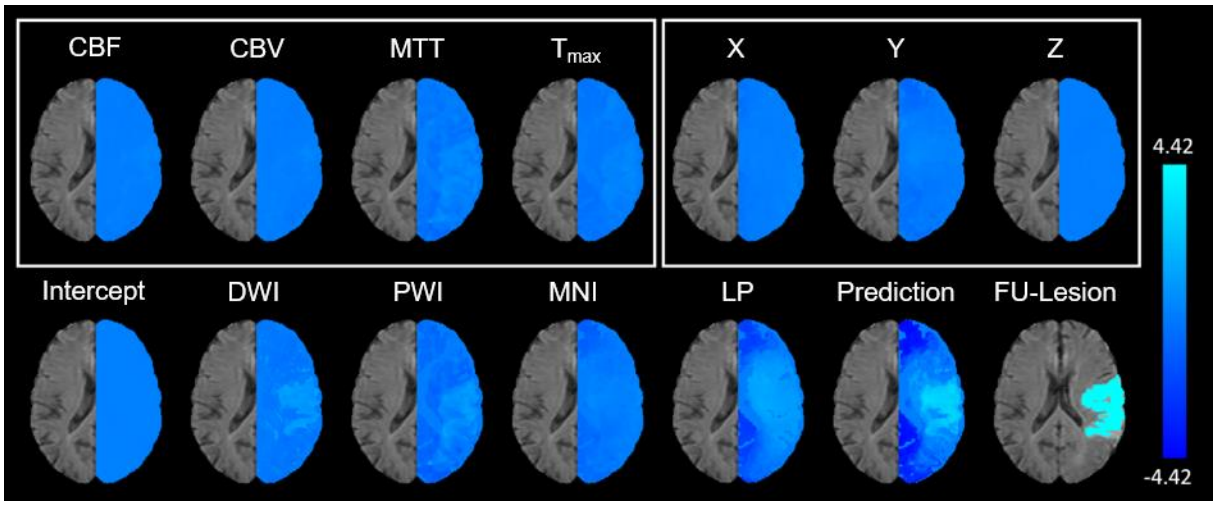

Abbildung 23: Shapley-Wert-Zerlegung: Shapley-Wert-Zerlegung der Vorhersagen des besten XGB-Modells mit den Features DWI (ADC), PWI, MNI und LP (Setting 6). Für jedes Voxel entspricht die Summe des Biasterms und der Shapley-Werte für DWI, PWI, MNI und LP der Vorhersage für die Follow-up-Läsion auf der Logit-Skala (untere Zeile). Die Shapley-Werte für PWI und MNI sind selbst Summen der Shapley-Werte von CBF, CBV, MTT, T_{max} (PWI) und X, Y, Z (MNI; obere Reihe). Eine weitere Addition der Shapley-Wertkarten für MNI und LP würde zu einer übergeordneten räumlichen Shapley-Wertkarte führen, die alle räumlichen Merkmale des Modells enthält. Solche Shapley-Wertkarten für die besten XGB-Modelle der anderen Featurekombination finden sich in Abbildung 24 (Quelle: Grosser et al., 2020a).

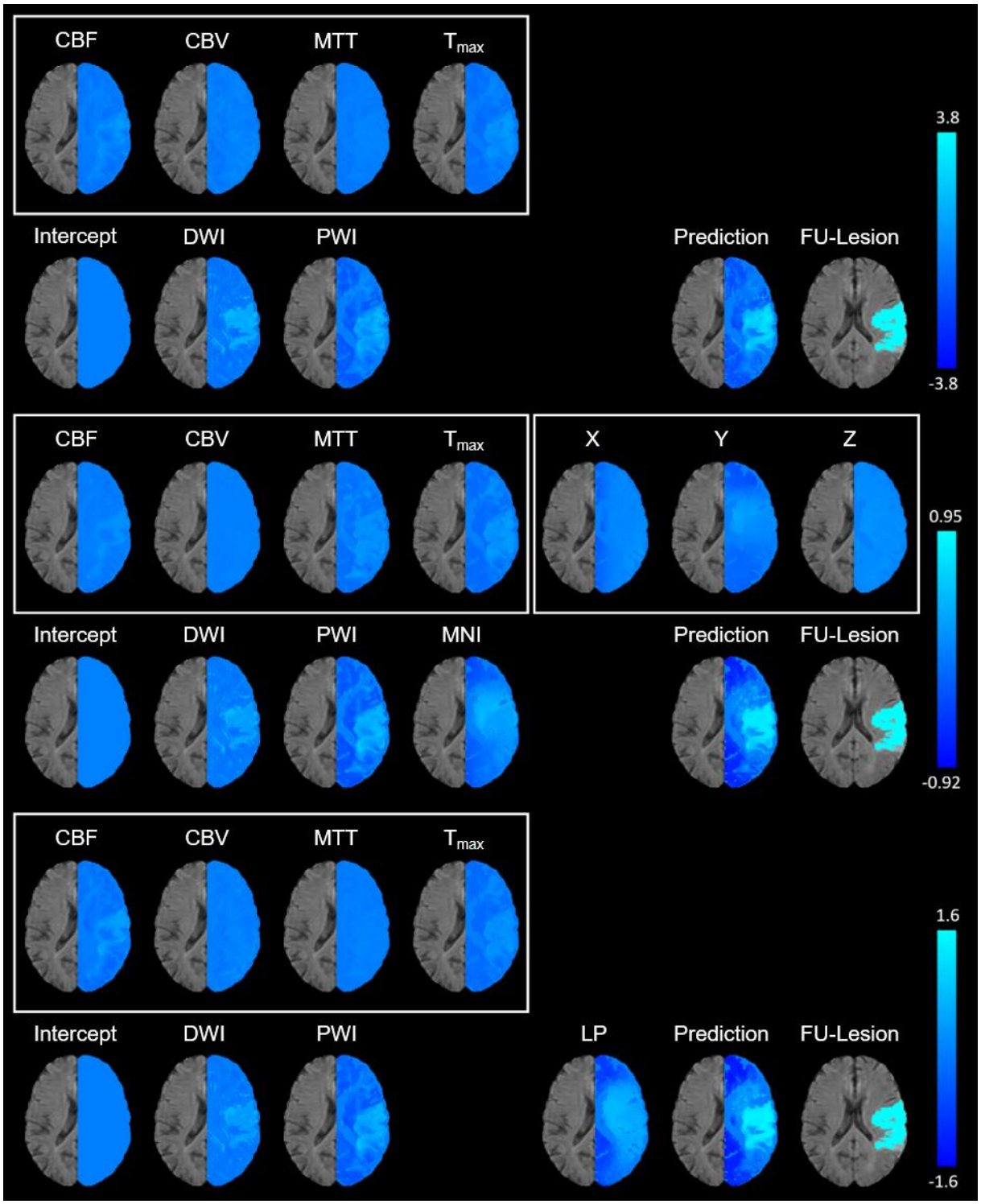

Abbildung 24: Shapley-Wert-Zerlegungen: Shapley-Wert-Zerlegungskarten der Vorhersagen der besten XGBoost-Modelle unter Einbezug von DWI und PWI (Setting 7; Zeilen 1–2), DWI, PWI und MNI-Koordinaten (Setting 3; Zeilen 3–4), DWI (ADC), PWI und LP (Setting 7; Zeilen 5–6) für einen ausgewählten Patienten (Quelle: Grosser et al., 2020a).

Permutation Importance

In den vorherigen beiden Unterabschnitten wurde zunächst untersucht, welche Gain-Werte die einzelnen Features prozentual während des Trainings der besten XGBoost-Modelle erzielten. Anschließend wurden die Featurebeiträge für die Vorhersagen mittels Shapley-Werten berechnet. Um den Einfluss der Features auf die Zielmetriken (Dice-Koeffizienten und ROC AUC) während der Kreuzvalidierung zu bestimmen, wurde jedes Feature des Modells (auf den Testdaten) zufällig permutiert. Dadurch wurde das betreffende Feature für die Vorhersage uninformativ, sodass sich dessen Permutation Importance aus der Differenz des Modells in der jeweiligen Metrik zu einem Modell mit vollständigen Informationen ergibt (siehe Abbildung 25).

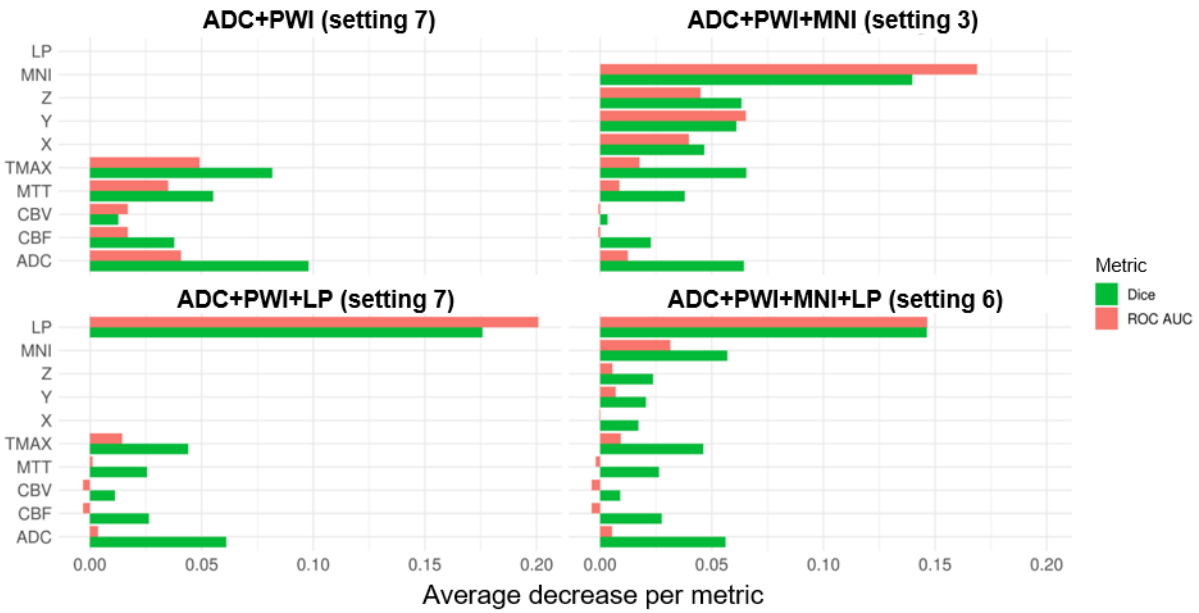

Abbildung 25: Permutation Importance pro Feature: Durchschnittliche absolute Abnahme der Dice- und ROC-AUC-Metriken für zufällige Permutationen jedes Features. Für die Berechnung der Dice-Koeffizienten wurde derselbe Ansatz zur Bestimmung des optimalen Schwellwerts verwendet wie für die anderen Berechnungen in dieser Arbeit. X-, y- und z-MNI-Koordinaten wurden gemeinsam permutiert, da eine Permutation von nur einer MNI-Achse zu Koordinaten führen kann, die vom Modell während des Trainings nicht gesehen wurden (Quelle: Grosser et al., 2020a).

Rangfolge

Zusammenfassend ergibt sich aus der Auswertung, dass das räumliche LP-Feature den höchsten Ranking-Score (0,975) erreichte. Damit war es das einflussreichste Feature im DWI-PWI-LP- und DWI-PWI-MNI-LP-Modell. Lediglich beim DWI-PWI-LP-Modell wurde das LP-Feature bzgl. des Gains von T_{max} übertroffen. Die MNI-Koordinaten (0,896) steuerten den größten Beitrag zur Optimierung des DWI-PWI-MNI-Modells und den dritthöchsten zum DWI-PWI-MNI-LP-Modell bei, wo sie – in Bezug auf den durchschnittlichen Gain – hinter LP und T_{max} lagen.

T_{max} hatte den dritthöchsten durchschnittlichen Ranking-Score und erreichte die besten Ergebnisse von allen PWI-Features. Diese waren meist in der Reihenfolge T_{max} (0,785) – MTT (0,442) – CBF (0,215) – CBV (0) angeordnet; mit Ausnahme des Gain-Rankings des DWI-PWI-Modells, wo T_{max} hinter den Ergebnissen von MTT rangierte und des DWI-PWI-MNI-LP-Modells, wo CBF vor MTT platziert ist.

Den vierthöchsten Durchschnittsrang erreichte der ADC (0,623), welcher stets vor CBF und CBV lag. Permutationen von ADC führten zur zweithöchsten Abnahme des Dice-Koeffizienten nach den räumlichen Merkmalen bzw. im Falle des DWI-PWI-MNI-Modells nach T_{max}. Darüber hinaus leistete der ADC den drittgrößten Beitrag in Bezug auf die ROC AUC Permutation Importance und die Shapley-Werte, gefolgt von den räumlichen Merkmalen und T_{max}.

MTT war das einzige Merkmal neben den räumlichen und T_{max}, welches teilweise höher rangierte als ADC (in den Gain-Rankings der DWI-PWI- und DWI-PWI-MNI-Modelle).

4.4 Vergleich zwischen lokalem Ansatz und globalem Ansatz mit räumlichen Features

Die Trainingsdaten der globalen Modelle aus beiden Ansätzen unterscheiden sich darin, dass sie für den ersten Ansatz stratifiziert gesampelt wurden, wohingegen die Voxel beider Outcomeklassen beim zweiten Datensatz ausbalanciert waren, indem die Nichtläsionsvoxel auf die Anzahl der Läsionsvoxel (pro Patient) heruntergesampelt wurden. Diesbezüglich konnten die Modelle ohne räumliche Informationen miteinander verglichen werden, wobei für das Random-Forest-Modell beim zweiten Ansatz eine hochsignifikante Verbesserung von 0,04 bzgl. der ROC AUC gezeigt werden konnte.

Zugleich war der Dice-Koeffizient bei diesem Modell um 0,02 höher. Keine relevanten Unterschiede zeigten sich hingegen für die beiden logistischen Regressionsmodelle.

Die strukturierten Hypothesentests für den Vergleich der lokalen und hybriden Modellierung mit der um räumliche Informationen erweiterten globalen Modellierung sind in Tabelle A6 im Anhang zu finden. Die lokale Modellierung erwies sich für die logistische Regression vor allem gegenüber der Verwendung von MNI-Koordinaten im globalen Ansatz als vorteilhaft: In diesem Fall ließ sich ein hochsignifikanter Unterschied von 0,03 bzgl. der ROC AUC feststellen sowie eine hohe, wenn auch nicht signifikante Effektstärke von 0,05 bzgl. des Dice-Koeffizienten. Beim Einbezug des LP-Features lieferte die globale Modellierung einen höheren Dice-Koeffizienten und eine niedrigere ROC AUC als die lokale Variante. Signifikante Unterschiede zeigten sich jedoch nicht.

Ein ähnliches Muster ergab sich beim Vergleich für die hybride logistische Regression: Diese lieferte eine um 0,05 höhere ROC AUC und einen um 0,06 höheren Dice-Koeffizienten als das globale Modell mit den MNI-Koordinaten als einzigem räumlichen Feature. Geringer fällt der Unterschied im Vergleich mit der globalen Modellierung aus, wenn diese das LP-Feature beinhaltet: In diesem Fall weist die hybride Modellierung jeweils einen um 0,03 höheren Dice-Koeffizienten und eine nur geringfügig niedrigere ROC AUC auf.

Der globale Random Forest mit räumlichen Features erreichte in beiden Metriken und für alle Featurekombinationen höhere Werte als der lokale und der hybride Random Forest.

Gegenüber dem lokalen Random Forest zeigten sich in Bezug auf die ROC AUC durchgängig hochsignifikante Verbesserungen von durchschnittlich 0,04. Für das Modell mit den MNI-Koordinaten als einzigem räumlichen Feature betrug die Differenz sogar 0,05. Dieses Modell zeigte mit 0,07 ebenfalls die größte Effektstärke bzgl. des Dice-Koeffizienten. Diesbezüglich führten allerdings auch die anderen Modelle mit je 0,06 zu erheblichen Verbesserungen in der Vorhersagequalität.

Für den hybriden Random Forest war der Unterschied zu einer globalen Modellierung etwas geringer. Einzig bei der ROC AUC des RF-Modells mit MNI-Koordinaten zeigte sich ein signifikanter Nachteil. Die Verbesserungen durch die globale Modellierung reichten hier von 0,02 (LP) bis 0,03 (MNI). Die positiven Effekte der räumlichen Features auf den Dice-Koeffizienten betrugen 0,02 (LP) und 0,03 (MNI).

5 Diskussion

5.1 Lokaler Ansatz

5.1.1 Vergleich der verschiedenen Modellvarianten

Nach den Berechnungen in dieser Studie erweisen sich lokal trainierte Machine-Learning-Modelle als vorteilhaft für die Läsionsprognose.

Verglichen wurden hierbei der lokale Ansatz, der mehrere lokale Modelle kombiniert (und durch das globale Modell ergänzt wird), mit einer globalen logistischen Regression und einem globalen Random-Forest-Modell sowie mit einem hybriden Ansatz, der sich aus dem Durchschnitt von lokalem und globalem Ansatz zusammensetzt. Für beide Algorithmen erzielte der lokale Ansatz bessere Ergebnisse in der Prognose der tatsächlichen Läsionen als der globale Ansatz. Der lokale Ansatz wurde allerdings noch vom hybriden Ansatz übertroffen. Der globale Random Forest führte zu einem etwas höheren Dice-Koeffizienten als der lokale Random Forest. Mit dem globalen Ansatz zeigten sich für Random Forest ebenfalls konsistente, jedoch geringfügige Verbesserungen gegenüber den beiden anderen Ansätzen in Bezug auf die Spezifität.

5.1.2 Erweiterung des lokalen Ansatzes

Als eine der wesentlichen Limitationen des vorgestellten lokalen Ansatzes stellte sich heraus, dass trotz der großen Datenbasis dieser Arbeit kein lokales Modell für 28 % der Gehirngewebepositionen im MNI-Raum trainiert werden konnte. Es stellt sich deshalb die Frage, wie gut der lokale Ansatz ohne die Erweiterung durch den globalen Ansatz gewesen wäre, obwohl an Positionen, an denen lokale Modelle wegen zu wenig Läsionsvoxeln nicht trainiert werden konnten, tendenziell eher leichter vorhersagbare Nichtläsionsvoxel vorkommen.

Für die Erweiterung des lokalen Ansatzes durch ein globales Modell wurde das globale Modell unter möglichst ähnlichen Bedingungen wie die lokalen Modelle trainiert. Dazu wurden für das Training der globalen Modelle die zufällig gezogenen Datensätze nach dem Gewebeoutcome stratifiziert und die Trainingsdaten der speziell für diesen Ansatz trainierten globalen Modelle wurden bewusst nicht ausbalanciert. Andernfalls könnten die Verteilungen der Läsions- und Nichtläsionsvorhersagen der globalen Mo-

delle zu stark von denjenigen der lokalen Modelle abweichen, sodass keine Kombination der beiden Ansätze (ohne eine geeignete Normalisierung) sinnvoll gewesen wäre.

5.1.3 Anreicherung der Trainingsdaten

Durch die Anreicherungsschritte der lokalen Trainingsdaten gelang es, dass der lokale Ansatz statt an 32,27 % an 72,33 % der Gewebepositionen trainiert werden konnte. Bereits das Zusammenlegen der Trainingsdatensätze aus unterschiedlichen Gehirnhälften ermöglichte ein Training der lokalen Modelle an 46,65 % der Positionen. Allein das Anreichern von maximal zwei Voxeln innerhalb eines Suchradius erhöhte den Grad der Abdeckung auf 59,58 % der Positionen.

Dabei wurde der Anreicherungsradius für die lokalen Trainingsdaten bewusst geringgehalten, um zu gewährleisten, dass die Grundvoraussetzung der Lokalität beibehalten wird. Voraussetzung für das Training eines lokalen Modells bestand in der Mindestanzahl von drei Voxeln pro Outcome, sodass gleichzeitig genügend Variabilität in den Trainingsdaten der lokalen Modelle vorhanden ist und genügend Modelle für den lokalen Ansatz trainiert werden können. Es ist eine offene Frage, ob die Mindestanzahl von drei Voxeln hier den besten Trade-off zwischen der Anzahl an trainierbaren Modellen und der Prognosegüte darstellt.

Die Anzahl der lokalen Modelle lässt sich dadurch erhöhen, dass Änderungen an den einzelnen Anreicherungsschritten vorgenommen werden. So kann beispielsweise der Anreicherungsradius erhöht werden oder die restriktive Anreicherung um maximal zwei Voxel pro Patienten für ein lokales Modell aufgehoben werden. Zudem besteht die Möglichkeit, einen größeren Datensatz zu verwenden. Letzterer sollte vor allem zusätzlich Infarkte im hinteren Stromgebiet beinhalten, da diese und lakunäre Infarkte bei der I-KNOW-Studie ausgeschlossen wurden.

Für die logistische Regression gilt, dass die Anzahl der Läsionen mit der Summe der Prädiktionen (auf den Trainingsdaten) übereinstimmt [Hosmer et al., 2013]. So können höhere Anteile an Läsionsvoxeln in den Trainingsdaten schnell zu durchschnittlich höheren Prognosen führen. Bei ungleichmäßig erhöhtem Läsionsaufkommen in den Trainingsdaten der lokalen Modelle lassen sich deren Vorhersagen deshalb schwieriger miteinander vergleichen, sodass dies auch die Bestimmung eines allgemeinen Schwellwerts für alle lokalen Modelle erschwert.

Alternativ ließe sich ein zusätzlicher Normalisierungsschritt oder verschiedene Schwellwerte für Gruppierungen von lokalen Modellen einführen, z. B. nach Gehirnregion oder nach Ähnlichkeit der Trainingsdaten. Dass die lokalen Trainingsdatensätze an sich kein konstantes Verhältnis bzgl. der beiden Outcomeklassen aufweisen, liegt in der Natur des Problems. Dies wirft die komplexe Frage auf, ob und für welche Algorithmen es vorteilhaft wäre, die Verteilungen im Rahmen eines weiteren Anreicherungsschrittes bewusst anzunähern. Dass dies für den globalen Ansatz sinnvoll ist, wurde bereits in [Jonsdottir et al., 2009] gezeigt. Für den lokalen Ansatz könnten ausgewogene lokale Trainingsdatensätze z. B. durch einen dynamischen Anreicherungsradius erzielt werden, sodass stets zwei Voxel unterschiedlicher Outcomes eines Patienten in jedes lokale Modell einfließen.

Alternative Modellierungen der Lokalität wären z. B. die Verwendung von Gehirnregionen und der Gewebeart zur Auswahl der Trainingsdaten. Auf Basis dieser Informationen könnten die Trainingsdaten eines lokalen Modells auf ausschließlich Voxel derselben Gehirnregion und/oder (der wahrscheinlicheren) Gewebeart der Modellposition limitiert werden. Beide Informationen sind wie in [Winder et al., 2019] über den MNI-Atlas abzurufen. Bei einer radialen Anreicherung von Voxeln (wie in dieser Arbeit) können die Trainingsvoxel je nach Abstand zur Position des Modells gewichtet werden. Während des Trainings werden die Datensätze invers zu ihrer Entfernung priorisiert, sodass weit entfernte Voxelpositionen ein geringeres Gewicht erhalten als lokale Informationen. Auf Basis dieser Gewichtung ist es daher möglich mehr Trainingsdaten innerhalb eines größeren Radius anzureichern, ohne die Grundvoraussetzung der Lokalität bei diesem Ansatz zu verlieren.

5.1.4 Skalierbarkeit

Das Training einzelner lokaler Modelle vollzieht sich komplett unabhängig voneinander, sodass die Modelle parallel trainiert werden können. Da beim (reinen) lokalen Ansatz die Anzahl der Trainingsdaten eines lokalen Modells lediglich der Patientenanzahl entspricht, skaliert dieser Ansatz für eine beliebig große Anzahl an Patienten. Die einzige Einschränkung der Unabhängigkeit zwischen den lokalen Modellen besteht in der leichten Überlappung der einzelnen Trainingsdatensätze durch die Anreicherungsschritte. Rechnerisch stellt dies jedoch kein Problem dar.

Sofern die Modelle auf eine geeignete Weise kombiniert werden können, ist für die einzelnen lokalen Modelle auch der Einsatz von anderen komplexeren Machine-

Learning-Modellen bei einer größeren Anzahl an Trainingsdaten denkbar. Die logistische Regression und der Random Forest wurden für diesen lokalen Ansatz gewählt, da sie vergleichsweise recheneffizient und robust (in Bezug auf eine Überanpassung an die Trainingsdaten) sind. Im Falle ausreichend vieler heterogener Patientendatensätze sollte zusätzlich eine Lernkurve erstellt werden, um die Auswirkungen größerer Trainingsdatensätze auf die Performance der lokalen Modelle besser abschätzen zu können.

5.1.5 MNI-Registrierung

Eine wichtige Voraussetzung für den lokalen Ansatz ist die Registrierung der Daten auf den MNI-Raum, welche allerdings zusätzliche Rechenkosten verursacht. Dazu wurden zunächst alle in dieser Arbeit eingeschlossenen Datensätze, die die anderen Bildverarbeitungsschritte erfolgreich durchlaufen haben, mithilfe von ANTs auf den MNI-Raum registriert.

Für die Auswahl des MNIs braucht es einen symmetrischen Atlas, damit sich die Symmetrie für den zweiten Anreicherungsschritt der lokalen Modelle nutzen lässt. Die Wahl fiel auf diesen MNI-Atlas, da dort ebenfalls eine Einteilung der Gehirnregionen vorhanden war, sodass die Koeffizienten der lokalen logistischen Regression für die einzelnen Gehirnregionen verglichen werden konnten. Die Auflösung dieses MNIs war mit $1 \times 1 \times 1\ \mathrm{mm}^3$ wesentlich höher als die der verwendeten MRT-Sequenzen. Zunächst sollte die Qualität der Registrierungen bei der Wahl eines anderen MNI-Raums kritisch geprüft werden, da die Auflösung der einzelnen MRT-Sequenzen jedoch stark nach Sequenz und Koordinatenachse variiert und in einigen Fällen sogar über der des MNIs liegt. So reichte beispielsweise die räumliche Auflösung der FLAIR-Schichten von $0{,}45 \times 0{,}45\ \mathrm{mm}^2$ bis $1 \times 1\ \mathrm{mm}^2$ bei einer Schichtdicke von 6 mm bis 7 mm.

Allgemein wäre ein MNI mit niedrigerer Auflösung vorteilhafter gewesen, um die Anzahl der MNI-Positionen und damit die Gesamtdatenmenge etwas zu reduzieren. Dadurch hätten in kürzerer Zeit wesentlich mehr Optimierungen in Form von verschiedenen Anreicherungsdesigns getestet werden können, wie zuvor beschrieben. Eine Alternative wäre es auch lediglich die Koordinaten aus der MNI-Registrierung zu verwenden, jedoch die Trainingsdaten in ihrem Ursprungsraum zu belassen und die Zuordnung der lokalen Trainingsdaten entsprechend anzupassen. Hiervon wurde jedoch

für eine konsistente Vergleichbarkeit aller Ansätze abgesehen, da die Auflösungen der rohen Bilddatensätze zwischen Patienten variieren.

Obwohl durch den verwendeten MNI die Auflösung um einen Faktor von etwa 20 gestiegen ist, prognostiziert die lokale logistische Regression das Gewebeoutcome nicht wesentlich langsamer als die globale logistische Regression. Dies liegt daran, dass die lokalen Modelle im Wesentlichen durch eine Koeffizientenmatrix repräsentiert werden können, sodass der lineare Vorhersageteil für einen Patienten auf einer simplen Matrixmultiplikation basiert und die finale Prognose anschließend mit der logistischen Funktion erhalten wird. Im Falle des Random Forests ist bereits der globale Ansatz mit 24,2 Sekunden für die Vorhersage bei einem neuen Patienten relativ langsam, allerdings noch ausreichend schnell im Gegensatz zum lokalen Random-Forest-Ansatz, der mit seiner derzeitigen Implementierung mit 705 Sekunden noch weitaus mehr Zeit benötigt. Hier kann die Möglichkeit der Parallelisierung des lokalen Ansatzes auch auf die Anwendung übertragen werden, sodass die für einen Rechenkern angegebene Zeit durch eine Parallelisierung über zusätzliche Kerne minimiert werden kann.

5.1.6 Implementierung weiterer Features

Die Ergebnisse dieses Ansatzes deuten darauf hin, dass räumliche Informationen die Vorhersage des Gewebeoutcomes beim ischämischen Schlaganfall verbessern. Deshalb sollten weitere (räumliche) Informationen in den lokalen Ansatz einbezogen werden. Je nach Anreicherungsschritten und verfügbaren lokalen Trainingsdaten lassen sich unterschiedlich viele Features integrieren.

In erster Linie muss die Lage des Voxels im Hinblick auf das akute Infarktgeschehen modelliert werden. Aufgrund der Knappheit der Trainingsdaten bietet sich hierzu der bereits von [Winder et al., 2019] verwendete Abstand zum Infarktkern sowie eine sorgfältige Auswahl an lokalen Umgebungsstatistiken analog zu [McKinley et al., 2016] an, anstelle der patchbasierten Nachbarschaftsmodellierung von [Benzakoun et al., 2021], bei der eine Vielzahl an Features ins Modell eingebaut werden muss.

Darüber hinaus wurde der lokale Ansatz im Hinblick besserer Integrationsmöglichkeiten von Patientenfeatures entwickelt. Hierzu zählt der Rekanalisationsstatus auf Basis des TICI-Scores und ggf. die Zeit bis zur Rekanalisation. Ein Vorteil des Ansatzes ist, dass aufgrund der unterschiedlichen Modellpositionen keine räumlichen Interaktionsterme in der Modellierung benötigt werden. Allerdings sollte eine Interaktion aus

DWI- und akuter FLAIR-Sequenz verwendet werden, um ein akutes DWI-FLAIR-Mismatch zu modellieren, was bei gelungener Rekanalisation die Chance auf ein gutes Voxeloutcome verbessert. Zusätzlich sollten Informationen über die Verschlusslage ins Modell einfließen, um zu erlernen für welche Verschlussorte die Modellposition Teil des Versorgungsgebiets ist. Für die hybriden Modelle sollte zusätzlich getestet werden, ob sich für Patienten mit erfolgreicher Rekanalisation eine höhere Gewichtung der lokalen Modelle – analog zur Wahrscheinlichkeitskarte in [Shen und Duong, 2008] – als vorteilhaft erweist.

Eine Anreicherung weiterer Patientenfeatures wie z. B. das Alter, das Geschlecht, die NIHSS oder die Zeit seit Infarktbeginn erscheint erst sinnvoll, wenn mehr Patientendatensätze für das Training zur Verfügung stehen. Zusätzlich kann die Gehirnhälfte als binäres Feature verwendet werden. Eine alternative Modellierung pro Gehirnhälfte würde die Annahme der Symmetrie des Läsionsrisikos zwar auflösen, allerdings die Menge der Trainingsdaten für alle lokalen Modelle etwa halbieren.

Bei Designs zur Vergrößerung der lokalen Trainingsdaten wie beispielsweise einem größeren Anreicherungsradius könnte es sinnvoll sein, die Gewebeart (Wahrscheinlichkeitswert für graue Substanz) als Feature in die Modellierung einfließen zu lassen. Bislang wurden Unterschiede zwischen grauer und weißer Hirnsubstanz im lokalen Ansatz nur implizit modelliert. Aufgrund der Anreicherungsmethode, die Voxel innerhalb eines kleinen Anreicherungsradius in die Trainingsdaten aufzunehmen, kann es jedoch leicht passieren, dass mehrere Voxel beider Gewebetypen in denselben lokalen Trainingsdaten vorkommen. Bei einer hybriden Modellierung sollte diese Information ebenfalls ins globale Modell integriert werden.

5.1.7 Zwischenfazit

Insgesamt zeigte sich eine Verbesserung der Vorhersagen des Gewebeoutcomes für die logistische Regression und den Random Forest mit dem lokalen und hybriden Ansatz. Die Vorhersageergebnisse sind aufgrund unterschiedlicher Datensätze nicht direkt mit denen aus der Literatur vergleichbar (vgl. Kapitel 5.4). Sie scheinen vor allem für den hybriden Ansatz in einem ähnlichen Rahmen wie die meisten aktuellen Shallow-Learning- und Deep-Learning-Methoden zu liegen, weisen jedoch klare Unterschiede zu absoluten State-of-the-Art-Ergebnissen wie [Benzakoun et al., 2021] oder [Yu et al., 2020] auf, was durch die hier vernachlässigte Modellierung von Voxelnachbarschaften erklärt werden kann.

Da für die vorliegende Studie nicht genug Patienten für eine „reine“ lokale Modellierung zur Verfügung standen, musste der lokale Ansatz mithilfe zweier Anreicherungsschritte und über ein globales Modell erweitert werden. Der Ansatz ist allerdings einfach auf eine größere Anzahl von Patienten skalierbar, wodurch er (bei einer heterogeneren Infarktverteilung) an mehr Positionen trainierbar wird und eine weitere Verbesserung gegenüber dem globalen Ansatz zu erwarten ist. Im Gegensatz zu anderen Arbeiten, die eine komplexe Modellierung der Voxelnachbarschaft enthalten wie [Benzakoun et al., 2021] oder tiefen neuronalen Netzen, müssen die Daten beim lokalen Ansatz nicht gesampelt werden, da sich die Trainingszeit aufgrund der Parallelisierbarkeit auch mit steigender Patientenzahl nicht sonderlich erhöht.

Neben der Patientenzahl spielt die Anreicherung der lokalen Trainingsdaten eine entscheidende Rolle. Für ausbalanciertere lokale Trainingsdatensätze scheint ein dynamischer Anreicherungsradius eine attraktive Alternative. Durch eine zusätzliche Restriktion nur Voxel aus der Atlasregion und des vorherrschenden Gewebetyps der Modellposition zu wählen sowie eine inverse Gewichtung zum Abstand der Trainingsvoxel zur Modellposition vorzunehmen könnte die Grundvoraussetzung der Lokalität des Ansatzes weiterhin gewährleistet werden.

Da räumliche Informationen im Ansatz bereits enthalten sind sollte bei der Integration weiterer Features zunächst der Fokus auf der Voxelnachbarschaft und dem Rekanalisationsstatus liegen. Daneben können bei steigender Datenmenge auch weitere Patientenfeatures wie u. a. in [Winder et al., 2019] sowie zusätzliche Biomarker wie in [Livne et al., 2018] einbezogen werden. Eine Methode, um relative Lageinformationen zu modellieren, ist die Nutzung des Abstandes zum Infarktkern wie in [Winder et al., 2019]. Verglichen mit der patchbasierten Nachbarschaftsmodellierung in [Scalzo et al., 2012] und [Benzakoun et al., 2021] ist der Vorteil der Abstandsinformation, dass diese in nur einem Feature kodiert wird. Ein interessantes weiteres Feature wäre die genauere Lokalisation des Verschlusses, weil damit erlernt werden könnte, ob die Position innerhalb des spezifischen Versorgungsgebietes liegt.

Nach der erfolgreichen Implementierung und Validierung des lokalen Ansatzes kann in zukünftigen Modellierungen für die Integration patientenspezifischer Features auf ihn zurückgegriffen werden, ohne dass dies zu einem systematischen Bias führt. Da der lokale Ansatz dennoch gegenüber einer globalen Modellierung einen Nachteil bzgl. der Trainingsdatenmenge aufweist wurden diese beiden Ansätze in dieser Arbeit zu einem hybriden Ansatz verschmolzen, der die Vorteile beider Ansätze beinhaltet.

Zukünftig können über die lokale Modellierung räumliche Informationen und Patientenfeatures in den hybriden einfließen und globale Zusammenhänge aus dem globalen Ansatz. Dabei weist der hybride Ansatz vor allem glattere Vorhersagen als der lokale Ansatz auf, sodass die resultierenden Vorhersagekarten in der Praxis auch einfacher zu bewerten sind.

5.2 Integration von räumlichen Features in den globalen Ansatz

5.2.1 Effekte auf die Läsionsvorhersagen

Nach der Kontrastierung der unterschiedlichen Modellierungen zeigte sich, dass globale Machine-Learning-Modelle mit räumlichen Informationen das Gewebeoutcome für akute Schlaganfallpatienten besser vorhersagen als entsprechende Modelle ohne räumliche Informationen. Lediglich bei der globalen logistischen Regression verschlechterte sich der durchschnittliche Dice-Koeffizient durch den Einbezug von MNI-Koordinaten als alleinigem räumlichen Feature gegenüber der Variante ohne räumliche Informationen. Darüber hinaus erzielten die tree-basierten XGBoost- und Random-Forest-Modelle nahezu State-of-the-Art-Ergebnisse in Hinblick auf die ROC AUC und Dice-Werte, worin sie die logistische Regression deutlich übertrafen.

5.2.2 Räumliche Features in linearen Modellen

Hinsichtlich der heterogenen und komplexen Struktur des Gehirns ist die Annahme eines monotonen Zusammenhangs zwischen unterschiedlichen Aspekten der Gehirnphysiologie und dem räumlichen Infarktrisiko ungeeignet für die Modellierung. Deshalb bietet sich die Modellierung durch einen linearen Term in der logistischen Regression nur in Bezug auf die LP-Karten an, nicht jedoch für die kartesischen MNI-Koordinaten, weil die einzelnen Koordinaten-Achsen lediglich Geraden durch das Gehirn beschreiben.

Da in diese Arbeit keine Infarkte des hinteren Stromgebiets eingingen, ist eine leichte Korrelation zwischen den MNI-Koordinaten und den Läsionsgebieten aus den Trainingsdaten plausibel (siehe Abbildung 16). Es ist daher davon auszugehen, dass das Modell diesen Zusammenhang tendenziell berücksichtigt und in peripheren Regionen außerhalb der relevanten Versorgungsgebiete zumindest im Mittel einer Fehleinschätzung von gesundem Gewebe aufgrund variierender PWI-Werte leicht

entgegenwirkt. Dies zeigt sich in einer Zunahme der ROC AUC um 1,3 %. Gleichzeitig führt diese Modellierung tendenziell zu einer höheren Risikobewertung in Gehirnbereichen mit hohem Läsionsaufkommen. Hier kommt es stärker auf eine nichtlineare Differenzierung der Versorgungsgebiete an, wozu das Modell in dieser Form nicht in der Lage ist, sodass es zu Überschätzungen der Läsionsgebiete kommt, die sich mit einem deutlichen Rückgang des Dice-Koeffizienten um 2,5 % bemerkbar machen.

Da statistisch ein wesentlich stärkerer monotoner Zusammenhang zwischen dem LP-Feature und den relevanten Versorgungsgebieten zu erwarten ist, lässt sich dieses Feature deutlich erfolgreicher in das lineare Modell integrieren als die MNI-Koordinaten. Im Gegensatz zu den MNI-Koordinaten führte der Einbezug des LP-Features zu einem signifikanten Anstieg der ROC AUC (6,09 %) und einem zumindest unveränderten Dice-Koeffizienten. Das Vorhersagebeispiel in Abbildung 19 zeigt, dass die logistische Regression auf Basis des LP-Features alles außerhalb der Versorgungsgebiete scharf von den üblichen Risikoregionen abgrenzt, was die Verbesserung für die ROC AUC erklärt. Innerhalb der Versorgungsgebiete zeigten sich jedoch stets erhöhte Prognosescores. Zurückzuführen ist dies auf die lineare Modellierung des LP-Features und vor allem das Fehlen von Interaktionstermen zwischen dem LP-Feature und den akuten Bildgebungsparametern. Darüber hinaus wäre gerade für die Prädiktionen innerhalb der Versorgungsgebiete der Rekanalisationsstatus als Feature von Bedeutung. So ist das Modell durch den starken Effekt des LP-Features leicht verzerrt, was in diesen kritischen Bereichen auch bei weniger auffälligen DWI und PWI-Werten zu einer Fehlklassifikation des Gewebes und somit einer Überschätzung der Läsionen führen kann (siehe Abbildung 19).

Da die logistische Regression gegenüber dem LP-Feature keine zusätzlichen Informationen aus den MNI-Koordinaten lernen konnte, führte ein ergänzender Einbezug dieses Features zu keinen weiteren signifikanten Verbesserungen.

5.2.3 Räumliche Features in tree-basierten Modellen

Im Gegensatz zum linearen Modell sind die XGBoost- und Random-Forest-Modelle nach den Resultaten in der Lage, die MNI-Koordinaten direkt und in plausibler Form zu nutzen. Beide Modelle bestehen aus vielen iterativ (XGBoost) bzw. parallel (Random Forest) trainierten Entscheidungsbäumen, die ihre Features mehrmals splitten. Deshalb erweisen sich diese Ensemblemodelle als äußerst robust in der Nutzung von nichtlinearen Features wie den MNI-Koordinaten.

Beim Training eines einzelnen Entscheidungsbaums können ein oder mehrere Splits entlang einer oder mehrerer MNI-Achsen gemacht werden, sodass die Voxel bei jedem räumlichen Split anschaulich durch eine zweidimensionale Hyperebene im MNI-Raum unterteilt werden. Die Kombination aller Hyperebenen des Ensemble-Modells fragmentiert den gesamten MNI-Raum in kleine Blöcke. Da die einzelnen Bäume auf jeder Seite eines räumlichen Splits unterschiedliche Risikowahrscheinlichkeiten abbilden, ergibt sich für jeden der Blöcke eine andere Läsionsrisikoverteilung, die ansonsten von den anderen Features aus der Bildgebung (inkl. ihrer Interaktionsterme) abhängt. Aus der Kombination der einzelnen Blöcke entsteht somit eine räumliche Randverteilung der Läsionen für fixe Wertebereiche der anderen Features, die aus den Daten erlernt wurde. Die Granularität dieser Karte(n) ist zwar geringer als die des LP-Features. Letztlich ist sie damit jedoch auch weniger rauschanfällig und nur durch die räumliche Variation des Läsionsaufkommens in den Trainingsdaten und die Komplexität des Modells beschränkt, die von den Hyperparametereinstellungen und der Heterogenität der Trainingsdaten abhängen.

Anhand der ähnlichen Resultate für die XGBoost- und Random-Forest-Settings, die jeweils die MNI-Koordinaten oder das LP-Feature beinhalten, zeigt sich, dass beide Features (indirekt) vergleichbare räumliche Information beinhalten. Die gleichzeitige Hereinnahme beider räumlicher Features bewirkte demzufolge lediglich im Falle der MNI-Koordinaten beim Random Forest eine zwar signifikante jedoch relativ geringe Verbesserung der ROC AUC (< 1 %).

Im Gegensatz zur logistischen Regression kam es bei den tree-basierten Modellen seltener zu Über- oder Unterschätzungen von Infarkten inner- bzw. außerhalb der aus den räumlichen Features erlernten Versorgungsgebiete. Dies liegt sehr wahrscheinlich an der automatischen Identifikation der Interaktionen zwischen den räumlichen Features und Features aus der Bildgebung. So werden die Versorgungsgebiete zwar auch bei dieser Modellierung klar vom restlichen Hirnparenchym abgegrenzt, jedoch ist das Modell besser in der Lage innerhalb der gefährdeten Bereiche den Prognosescore weiter anhand der ADC- und PWI-Features für das tatsächliche Läsionsoutcome zu justieren.

Hinzu kommt, dass bei einigen der Patienten aus I-KNOW eine Rekanalisation herbeigeführt werden konnte. In Teilen der Versorgungsgebiete mit tendenziell besserer Kollateralversorgung fällt somit die Läsionshäufigkeit geringer aus als im Rest der Ver-

sorgungsgebiete, sodass dieser Aspekt ebenfalls in den räumlichen Features enthalten ist. Damit dienen die räumlichen Features als Indikator für solche Regionen.

Obwohl visuelle Prüfungen keine starken Überanpassungen in Gebieten mit hohen LP-Werten zeigten, kam es dort allerdings auch für die tree-basierten Modelle zu leichten Verzerrungen des Prognosescores und damit zu Überschätzungen der Infarkte. Diese Überanpassungen machten sich u. a. in der höheren Differenz zwischen der ROC AUC auf den Trainingsdaten und den Testdaten bei den tree-basierten Modellen bemerkbar. Dies liegt sehr wahrscheinlich daran, dass sich die Prognose des Gewebeschicksals für diese Bereiche vor allem mit dem rechtzeitigen Einsetzen einer Widerdurchblutung schlagartig verbessern kann. Der Rekanalisationsstatus wurde jedoch nicht in die Modellierung einbezogen, weil für die Patienten aus der I-KNOW-Studie der Zeitpunkt einer geglückten Rekanalisierung nicht bekannt ist, da dieses Ereignis erst in der Folgeuntersuchung festgestellt wurde.

5.2.4 Einfluss der Features in XGBoost-Modellen

Für den Gain, die Shapley-Werte und die Permutation Importance der untersuchten XGBoost-Modelle lieferten die räumlichen Features nahezu durchgehend die höchsten Beiträge und wiesen damit den deutlich größten Einfluss an den Modellen auf. Da in die I-KNOW-Studie keine Patienten mit Infarkten im hinteren Stromgebiet aufgenommen wurden, bewirken die räumlichen Features in tree-basierten Modellen zunächst eine klare Abgrenzung der verschiedenen Versorgungsgebiete.

Obwohl der Anteil des LP-Features dabei teilweise besonders herausragte und über dem der (aggregierten) MNI-Koordinaten lag, scheinen die Informationen vergleichbar, die in diesen beiden Features kodiert sind. Gestützt wird dies durch die ähnlichen quantitativen Ergebnisse der Modelle in Bezug auf die ROC AUC und Dice-Koeffizienten, die theoretischen Aspekte zur Wirkungsweise der räumlichen Features in tree-basierten Modellen (vgl. Abschnitt 5.2.3) sowie die visuell nahezu deckungsgleichen Shapley-Wert-Karten der MNI-Koordinaten und des LP-Features der DWI-PWI-MNI- und DWI-PWI-LP-Modelle (siehe Abbildung 24).

Abgesehen von den räumlichen Features war das T_{max}-Feature über alle Modelle hinweg das informativste Feature, gefolgt vom ADC und MTT. Während MTT besonders hohe Werte beim in der Trainingsphase gemessenen Gain aufweist, führt eine Permutation des ADC zu einem hohen Abstieg des Dice-Koeffizienten. Dieser Befund

deckt sich mit vielen Studien, in denen der ADC zur Bestimmung des Infarktkerns herangezogen wurde, während T_{max} und MTT größtenteils zur Bestimmung der Penumbra eines akuten Infarkts dienten. Insbesondere beobachteten [Livne et al., 2018] für ihr XGBoost-Modell ebenfalls den größten Einfluss für PWI-basierte Zeitkarten, gefolgt von Parametern aus der DWI-Bildgebung. Daher überrascht es kaum, dass diese Features einen hohen Beitrag zur Vorhersage leisten. Im Gegensatz dazu waren die CBF- und CBV-Features fast immer am wenigsten informativ, was sich mit der hohen Variation von CBF und CBV im Gehirngewebe (z. B. Unterschiede zwischen weißer und grauer Substanz) erklären lässt.

5.2.5 Integration weiterer Features

Sowohl die logistische Regression als auch die tree-basierten Modelle haben anhand der räumlichen Features gelernt, die relevanten Versorgungsgebiete für Verschlüsse im vorderen Stromgebiet vom Rest des Gehirnparenchyms abzugrenzen. Innerhalb der Versorgungsgebiete mangelt es der logistischen Regression zunächst an Interaktionstermen mit Features aus der akuten Bildgebung.

Am dringendsten fehlen allen Modellen aus dieser Arbeit Informationen über eine Rekanalisierung, deren Eintreten und Eintrittszeitpunkt erst im weiteren Verlauf bekannt sind. Für die Modellierung des Rekanalisationsstatus gibt es verschiedene Alternativen: Die sauberste scheint die Modellierung nach Rekanalisationsstatus wie bei [Winder et al., 2019] zu sein. Alternativ kann auch der Ansatz, der von [Pinto et al., 2018a] für neuronale Netze erprobt wurde, nämlich falsche Vorhersagen einzelner Klassen je nach Rekanalisationsstatus unterschiedlich zu bestrafen, auf die Zielfunktion von XGBoost übertragen werden. Dieser Ansatz hat den Vorteil, dass der Trainingsdatensatz nicht wie bei [Winder et al., 2019] zerteilt werden muss und stets auf der gesamten Datenbasis trainiert werden kann. Spätestens bei der Integration der Rekanalisationszeit stellt sich allerdings erneut die Frage, wie dieses Patientenfeature auf Voxelebene zu integrieren ist. [Benzakoun et al., 2021] verwenden den Rekanalisationsstatus als kategorielles Feature. Da sie ebenfalls die MNI-Koordinaten nutzen, wäre eine räumliche Auswertung der Shapley-Werte-Karten für die MNI-Koordinaten und den Rekanalisationsstatus interessant, um zu verstehen, ob durch dieses Feature in tree-basierten Modellen wirklich ein Bias entsteht, oder ob die Modelle lediglich innerhalb der Versorgungsgebiete, die über die MNI-Koordinaten gelernt werden, ein niedrigeres Läsionsrisiko (bei positivem Rekanalisationsergebnis) anzeigen. Abgese-

hen davon muss für diese Art der Features zunächst schlicht versucht werden, einer Überanpassung an die Trainingsdaten im Rahmen eines sorgfältigen Hyperparameter-Tunings während der Kreuzvalidierung entgegenzuwirken. Eine weitere Integrationsmöglichkeit besteht darin, etwas Random-Noise zur Rekanalisationszeit hinzuzufügen, sodass es tree-basierten Modellen unmöglich ist, Patienten über Splits entlang eines solchen Patientenfeatures eindeutig zu identifizieren. Alternativ besteht die Möglichkeit, Patientenfeatures innerhalb des lokalen Ansatzes zu modellieren und diesen mit dem globalen Ansatz – analog zum hybriden Modell (ggfs. mit optimierter Gewichtung) – zu kombinieren.

Um das Potential einer Rekanalisation besser zu beurteilen, werden weitere Marker für den Status der Kollateralen benötigt, da diese für ein langes Aufrechterhalten einer Penumbra überlebensentscheidend sind. Hierzu ist z. B. der Vorschlag von [Galinovic et al., 2018] aufzugreifen, der das Verhältnis von CBF zu T_{max} als Marker für den Kollateralstatus empfiehlt. Dieses sollte explizit als Feature kodiert werden, da keiner der in dieser Arbeit verwendeten Algorithmen automatisch Quotienten einzelner Features bildet. Weiterhin sollte auch die FLAIR-Sequenz genutzt werden, da sie zusammen mit dem ADC das DWI-FLAIR-Mismatch repräsentiert und somit einen Indikator für das Alter von Infarzierungen und das akute Infarktgeschehen liefert.

Zusätzlich können anatomische Marker aus einer Berechnung des LP-Features (vorzugsweise separat nach Verschlussort) für Patienten ohne Rekanalisation und mit Rekanalisation abgeleitet werden: Regionen mit niedriger Infarkthäufigkeit in beiden Karten sind offensichtlich außerhalb der Versorgungsgebiete der integrierten Verschlüsse. Regionen mit hohen Werten in beiden Karten sind besonders anfällig für Ischämien, da sie sowohl mit als auch ohne Rekanalisation ein schlechtes Voxeloutcome aufweisen. Bereiche, die für rekanalisierte Patienten eine deutlich geringere Läsionshäufigkeit aufweisen als für unverminderte Infarkte, profitieren in diesem Sinne am meisten von einer Rekanalisation, was sich in der Differenz der beiden Karten ausdrückt, die ein Marker für Regionen mit guter Kollateralenbildung darstellen könnte.

Demgegenüber können Nachbarschaftsinformationen, die vor allem eine kontinuierliche Ausbreitung des Infarktkerns modellieren, relativ zielgerichtet über den ursprünglich bei [Scalzo et al., 2012] und später von [Benzakoun et al., 2021] angewendeten patchbasierten Ansatz integriert werden. Da dieser Ansatz relativ viele Features benötigt, was die Nachbarschaftsgröße räumlich begrenzt, sollte außerhalb der direkten Nachbarschaft eine niedrigere Auflösung der Features gewählt werden.

Dazu können analog zu [McKinley et al., 2016] Statistiken über größere Voxelgebiete berechnet werden, sodass die Tree-Ensembles ähnlich wie CNN eine größere (hierarchische) Umgebung abdecken. Komplementär sollte der Abstand zum Infarktkern wie bei [Winder et al., 2019] in die Modelle einfließen, um das relative Risiko aufgrund der Lage zum Infarktkern besser zu beurteilen. Als Erweiterung sollte der Abstand entlang aller drei Koordinatenachsen als Feature kodiert werden, sodass auch die relative Lage zum Infarktkern in das Modell eingeht.

Nach den Ergebnissen dieser Studie trägt die Kenntnis des Versorgungsgebiets erheblich zur Eingrenzung des gefährdeten Gebietes bei. Insbesondere erlauben es die MNI-Koordinaten tree-basierten Modellen, die Informationen, die für die Verschlüsse im vorderen Stromgebiet in der LP-Karte kodiert sind, selbstständig aus den Daten zu lernen. Eine Limitierung dieser Modellierung ist, dass die so erlernten Versorgungsgebiete zum einen sehr allgemein gehalten sind und zum anderen nicht für Infarkte des hinteren Stromgebiets gelten. Eine individuellere Beurteilung könnte vorgenommen werden, wenn der Ort des Verschlusses in die Modellierung einfließt. Die tree-basierten Modelle könnten den Zusammenhang zwischen MNI-Koordinaten und Verschlussort als Interaktion erkennen, sodass Versorgungsgebiete individuell nach Ort des Verschlusses verstanden werden. Alternativ können solche Karten auch extern für jeden Verschlussort an Patienten ohne Rekanalisation erhoben werden und als eigenständiges Feature integriert werden. Im Falle der logistischen Regression müsste pro Karte ein Interaktionsterm mit dem Ort des Verschlusses spezifiziert werden.

Weiterhin lässt sich eine Differenzierung nach Gehirnhälfte vornehmen, deren Einfluss in dieser Arbeit, vor allem zur Umsetzung des lokalen Ansatzes, vernachlässigt wurde. Weitere räumliche Informationen wie der Gewebetyp und die Gehirnregion können analog zu [Winder et al., 2019] auf MNI-Basis modelliert werden.

Neben den typischen DWI- und PWI-Parametern können aus diesen Sequenzen weit mehr Informationen generiert werden als üblicherweise genutzt werden. So sollten beispielsweise rohe PWI-Karten (nach den üblichen Korrektur- und Standardisierungsschritten) auch in tree-basierten Modellen erprobt werden.

5.2.6 Zwischenfazit

Für alle drei Algorithmen aus dieser Arbeit zeigte sich eine signifikante Verbesserung in der Vorhersage des Gewebeoutcomes bei der Integration räumlicher Informationen.

Insbesondere die tree-basierten Modelle erwiesen sich als besonders robust in der Verarbeitung der räumlichen Features und erreichten deutlich bessere Ergebnisse als die logistische Regression. Dies liegt in erster Linie an ihren beiden Eigenschaften, nichtlineare Zusammenhänge und Interaktionen zwischen verschiedenen Features automatisch zu erlernen. Während der logistischen Regression vor allem eine Interaktion zwischen dem LP-Feature und den Features aus der akuten Bildgebung fehlte, erwiesen sich die MNI-Koordinaten als gänzlich ungeeignet für diesen Algorithmus.

Alle Algorithmen profitierten in erster Linie von den räumlichen Features, indem sie es erlaubten, Regionen mit niedrigem Infarktaufkommen hinsichtlich des Risikos scharf abzugrenzen von Gebieten, die bei Verschlüssen im vorderen Stromgebiet häufiger in Mitleidenschaft gezogenen werden. Einfacher ausgedrückt erlernten die Modelle die Versorgungsgebiete des vorderen Stromgebiets.

Innerhalb der Versorgungsgebiete weisen jedoch alle Modelle Schwächen bei einer schärferen Differenzierung zwischen zukünftigen Läsionen und gesundem Gewebe auf. Dies liegt daran, dass der I-KNOW-Datensatz Patienten mit und ohne erfolgreiche Rekanalisation enthält. Dementsprechend lassen sich die Ergebnisse aus dieser Studie ohne weiteres durch den Einbezug des Rekanalisationsstatus verbessern. Eine weitere Verbesserung sollte sich über den Einbezug von Nachbarschaftsinformationen realisieren lassen, die vor allem für tree-basierte Modelle über den patchbasierten Ansatz modelliert werden können.

Um eine weitere Differenzierung der Versorgungsgebiete auf Patientenebene zu ermöglichen, wäre der Einbezug des konkreten Verschlussortes eine Weiterentwicklung zu dieser Arbeit. So können z. B. spezifisch Läsionsverteilungen pro Verschlussort für Infarkte nichtrekanalisierter Patienten analog zum LP-Feature verwendet werden. Für die logistische Regression müsste ein Interaktionsterm zwischen Verschlussort und dem jeweiligen Versorgungsgebiet spezifiziert werden. Allergings würde es für tree-basierte Modelle wahrscheinlich schon ausreichen, den Verschlussort gemeinsam mit den MNI-Koordinaten in die Modellierung aufzunehmen, um festzustellen, ob ein Voxel zum Versorgungsgebiet des konkret verschlossenen Arterienzweigs gehört und aufgrund seiner Lage vom Infarkt betroffen sein könnte. Denn in dieser Arbeit wurde festgestellt, dass tree-basierte Modelle nahezu deckungsgleiche Informationen aus den MNI-Koordinaten und dem LP-Feature erlernen, was sich hier durch das automatische Erlernen von Interaktionen mit dem Verschlussort ausnutzen lässt, um spezifische Versorgungsgebiete zu lernen.

Es wurde bereits gezeigt, dass tree-basierte Modelle basierend auf Bildinformationen, räumlichen Features und Nachbarschaftsinformationen mit aktuellsten Deep-Learning-Modellen auf diesem Gebiet konkurrieren können [Benzakoun et al., 2021]. Besonders bei größeren Datenmengen lässt sich in zukünftigen Arbeiten herausfinden, inwieweit hier eine Verbesserung der Deep-Learning-Modelle zu erwarten ist. Wie z. B. [Winder et al., 2019] zeigte auch diese Studie, dass tree-basierte Ensemblemodelle der logistischen Regression bei der Vorhersage des Gewebeschicksals beim ischämischen Schlaganfall überlegen sind, weshalb sie in zukünftigen Arbeiten zumindest als Vergleichsmethode zum Einsatz kommen sollten. Die in dieser Arbeit verwendeten MNI-Koordinaten bieten eine einfache Lösung, räumliche Effekte in diesen Vergleichsmodellen zu berücksichtigen.

Bei der weiteren Entwicklung und der Etablierung von Modellen in klinischen Settings stellt sich die Frage, inwiefern die Erklärbarkeit von Modellen ausschlaggebend für deren Implementierung in der medizinischen Praxis ist. Insbesondere die in dieser Arbeit verwendeten Shapley-Werte bieten medizinischem Fachpersonal ein konsistentes Framework, um den Einfluss einzelner Features zu vergleichen und finale Vorhersagekarten transparent in die Einflüsse einzelner Features zu zerlegen und damit einen erklärbaren Zusammenhang zwischen den Daten und der Modellvorhersage zu liefern.

5.3 Vergleich beider Modellierungsansätze

5.3.1 Läsionsvorhersage

Insgesamt folgten aus der Integration von räumlichen Features in tree-basierte globale Modelle die besten Vorhersageergebnisse in dieser Arbeit bzgl. der ROC AUC und dem Dice-Koeffizienten. Hier erreichten die XGBoost-Modelle mit Werten von bis zu 0,89 bzw. 0,40 eine Steigerung von 0,02 bzw. 0,05 gegenüber der hybriden Modellierung aus dem ersten Forschungsansatz.

In allen Fällen profitierten die Modelle jedoch von einer Berücksichtigung der räumlichen Information: So erreichte selbst das beste globale XGBoost-Modell nur eine ROC AUC von 0,83 und einen Dice-Koeffizienten von 0,35, wenn ausschließlich Bilddaten zur Verfügung standen. Die globale logistische Regression kam in diesem Fall lediglich auf Werte von 0,81 bzw. 0,32.

Keine signifikanten Unterschiede ergaben sich für eine lokale, hybride und globale Modellierung (mit LP-Feature) in Hinblick auf die logistische Regression. Dennoch wies der hybride Ansatz mit einer Steigerung von 0,03 im Dice-Koeffizienten eine deutlich erhöhte Treffsicherheit für Läsionen gegenüber einer Integration des LP-Features in den globalen Ansatz auf. Nachdem der hybride Ansatz in Regionen mit geringem Infarktaufkommen durch den globalen Ansatz (ohne räumliche Informationen) ersetzt wurde, der keine Differenzierung solcher Bereiche zu den gefährdeten Versorgungsgebieten des vorderen Arteriensystems beinhaltet, sollte beim lokalen und hybriden Ansatz in Zukunft ein globales Modell, das räumliche Informationen beinhaltet, zur Erweiterung verwendet werden.

Für Random Forest war das globale Modell mit räumlichen Features in allen Fällen dem lokalen und hybriden Ansatz überlegen. Die größten Unterschiede zeigten sich in beiden Metriken für das Modell mit MNI-Koordinaten. Darüber hinaus sind – ähnlich wie für die logistische Regression – geringere Nachteile für den hybriden als für den lokalen Ansatz festzuhalten. Dabei ist zu bedenken, dass für die lokalen Modelle pro Position nur sehr wenige Datensätze zur Verfügung standen und die globalen tree-basierten Modelle besonders gut MNI-Koordinaten verarbeiten können. Vor diesem Hintergrund ist es bemerkenswert, dass der hybride Random Forest immerhin einen Dice-Koeffizienten von 0,35 aufweist und trotz signifikanten Unterschieden bzgl. der ROC AUC zu den Modellen mit MNI-Koordinaten nur relativ kleine negative Effektstärken von maximal -0,03 aufgetreten sind.

Die in anderen Arbeiten festgestellten Verbesserungen für Modelle, die auf ausbalancierten Datensätzen trainiert wurden wie u. a. [Winder et al., 2019], können in dieser Arbeit nur für Random Forest bestätigt werden. So wies der auf ausbalancierten Trainingsdaten trainierte Random Forest aus dem zweiten Forschungsansatz einen signifikanten Unterschied in der ROC AUC auf und eine (nicht signifikante) Verbesserung um 0,02 im Dice-Koeffizienten.[27]

5.3.2 Erklärbarkeit

Komplexere Modelle wie tree-basierte Ensembles oder neuronale Netze führen häufig zu besseren Ergebnissen in der Vorhersagequalität als einfachere Modelle wie die

[27] Hierzu ist anzumerken, dass in beiden Ansätzen unterschiedliche Random-Forest-Implementierungen verwendet wurden.

logistische Regression. Sie gelten jedoch aufgrund ihrer Komplexität als Blackbox, weshalb sie für ihre schwierige Interpretierbarkeit und damit fehlende Praxistauglichkeit oftmals kritisiert werden. In Anbetracht des Dilemmas zwischen Nutzen und Transparenz der Modelle wurden in dieser Arbeit zum ersten Mal Shapley-Werte erprobt, die als Framework zur Erklärung von Vorhersagen des Gewebeschicksals dienen können.

Shapley-Werte bieten für binäre Klassifikationsmodelle eine analoge Interpretation zum Produkt aus Koeffizienten und Feature im linearen Term einer logistischen Regression. Besonders effizient ist ihre Berechnung für tree-basierte Modelle wie z. B. XGBoost oder Random Forest [Lundberg et al., 2020]. Sie lassen sich jedoch für jedes Klassifikationsmodell berechnen.

Ein großer Vorteil von Shapley-Werten gegenüber anderen Interpretationstechniken wie z. B. local interpretable model-agnostic explanations kurz LIME [Ribeiro et al., 2016] ist ihre Konsistenzeigenschaft [Lundberg et al., 2019]. Während LIME für jede Vorhersage ein individuelles lineares Modell trainiert, berücksichtigen Shapley-Werte alle möglichen Kombinationen von Features. Hierdurch spiegeln Shapley-Werte unterschiedlich hohe Beiträge eines Features für verschiedene Vorhersagen konsistent wider. Für die Läsionsvorhersage spielt dies eine wichtige Rolle, da dadurch die Vorhersagen für alle Voxel eines Patienten im Zusammenspiel betrachtet werden. So zeigen Abbildung 23 und Abbildung 24 konsistente Shapley-Wert-Karten, die die XGBoost-Vorhersagen für einen Patienten in die Anteile der einzelnen Features zerlegen. Die Erstellung dieser Karten dauerte aufgrund der effizienten Berechnung von Shapley-Werten für tree-basierte Modelle weniger als eine Sekunde.

Ein weiterer Vorteil dieser Methodik liegt in der gemeinsamen Darstellung der PWI-Features und der MNI-Koordinaten, welche die Shapley-Werte-Karten in Abbildung 24 beinhalten. Der Einfluss dieser hierarchischen Features wurde sowohl auf Ebene der einzelnen Features als auch für mehrere Features gemeinsam als Summe der einzelnen Shapley-Werte dargestellt.[28] Vor allem für Modelle mit vielen Features wie die von [Benzakoun et al., 2021] trainierten tree-basierten Modelle, die insgesamt 453 Features beinhalten, kann die Additivität von Shapley-Werten genutzt werden, um die

[28] Die Visualisierung hierarchischer Entitäten auf unterschiedlichen Ebenen wird auch als Drilldown (Wechsel auf eine niedrigere Granularitätsstufe) bzw. Drillup (Wechsel auf eine höhere Granularitätsstufe) bezeichnet.

Komplexität der vielen – vermutlich größtenteils redundanten – Nachbarschafts- und kontralateralen Informationen bedarfsweise zusammenzufassen.[29]

Mithilfe der Shapley-Werte-Karten konnten in dieser Arbeit die Anteile der verschiedenen Features an der finalen Prädiktion auf ihren Beitrag überprüft werden. So konnte gezeigt werden, dass die Beiträge der MNI-Koordinaten und des LP-Features räumlich nahezu deckungsgleich sind (siehe Abbildung 24). Ebenfalls stimmt der Bereich der DWI-Läsionen mit der ADC-Shapley-Werte-Karte überein und liegt innerhalb des erhöhten Risikobereichs der PWI-Shapley-Werte-Karte (siehe Abbildung 23).

Im Falle von gänzlich fehlerhaft erlernten Zusammenhängen gestattet die reine Prognosekarte keine Möglichkeit, die konkrete Ursache festzustellen. Im Zweifelsfall können über Shapley-Werte-Karten und ähnliche Interpretationstechniken falsch erlernte Zusammenhänge entlarvt werden [Zihni et al., 2021].

Zur Bestimmung der Feature Importance für XGBoost-Modelle wurden in dieser Arbeit u. a. die durchschnittlichen Differenzen der Shapley-Werte zwischen Läsions- und Nichtläsionsvoxeln berechnet. Die Differenzen geben an, wie stark sich ein Feature im Modell auf die Differenzierbarkeit der verschiedenen Outcomes auswirkt (siehe Abbildung 22). Damit bietet sich die Differenz als eine besonders gut interpretierbare Alternative zum Gain und der Permutation Importance an. Die Rangfolge der Features nach diesem Maß erwies sich als erstaunlich konsistent zu den bereits etablierten Metriken. Es zeigte sich, dass die räumlichen Features (LP und MNI) am wichtigsten sind und die Versorgungsgebiete abgrenzen. Danach kamen T_{max} und MTT, die zur Abgrenzung der Durchblutungsstörung verwendet werden. Ebenso dann ADC, welches wiederum den Kerninfarkt innerhalb der Durchblutungsstörung abgrenzt (und damit ein besonders scharfes Kriterium für gefährdete Voxel darstellt). Erst danach kamen CBF und CBV, die allerdings aufgrund gewebeabhängiger Perfusionswerte stark variieren.

5.3.3 Limitationen

Eine zentrale Limitation des lokalen Ansatzes ist die niedrige Anzahl an Patientendaten, die für das Training zur Verfügung standen. Allerdings konnten die lokalen Trainingsdaten in vielen Fällen ausreichend mit weiteren Voxeln für das Training

[29] Die Additivität von Shapley-Werten ist gegeben, da diese im Vergleich zu einzelnen Features einheitlich skaliert sind und lediglich den Anteil eines Features an einer einzelnen Prädiktion beschreiben.

angereichert werden. Dennoch musste der lokale Ansatz an 28 % der MNI-Positionen durch ein globales Modell substituiert werden.

Vor diesem Hintergrund wurde bewusst der I-KNOW-Datensatz gewählt, der wesentlich mehr Patienten als die 43 frei verfügbaren Datensätze aus den ISLES-Wettbewerben umfasst. Zudem sind die Rohdaten der DWI- und PWI-Sequenzen für den ISLES-Datensatz nicht verfügbar. Da sie im Vergleich zu den Daten in dieser Arbeit unterschiedlich prozessiert wurden, was potenziell zu einem systematischen Bias führen könnte, wurden die ISLES-Daten in dieser Arbeit nicht eingeschlossen.

Für den globalen Ansatz mit räumlichen Features wurden ebenfalls ausschließlich die I-KNOW-Daten verwendet. Dies ermöglicht einen konsistenten Vergleich mit dem lokalen Ansatz, was Ziel dieser Arbeit war. Gleichzeitig erschwert dies jedoch die Kontrastierung mit Arbeiten, die auf den ISLES-Daten basieren.

Bildgebungsdaten unterscheiden sich durch den Akquirierungsprozess, die patientenabhängigen Eigenschaften und den Schlaganfall an sich. Zusätzlich wurde für einige Patienten aus den I-KNOW-Daten die erste Follow-up-Untersuchung erst nach sieben Tagen durchgeführt. In dieser Zeit kann sich die Läsionsgröße z. B. durch Änderungen der enthaltenen Wasseransammlung verändern. Die darauf basierenden Modelle können zu einer Überschätzung der tatsächlichen Läsionen neigen. In dieser Arbeit sind beim Prozessieren der Daten keine Schwellungen (oder Verschiebungen des Gehirns) aufgefallen.

Eine mangelnde Reproduzierbarkeit ist auch womöglich auf verschiedene Schritte in der Vorverarbeitungspipeline zurückzuführen, wie unterschiedliche Software oder Unterschiede in den Läsionssegmentierungen.

Die in dieser Arbeit verwendeten Daten wurden in verschiedenen Krankenhäusern nach demselben Studienprotokoll erhoben. Die angewendete Vorverarbeitungspipeline beinhaltet ausschließlich standardisierte Analysen und Bildsequenzen. Die Segmentierungen wurden von zwei Ratern in Übereinstimmung vorgenommen. Dadurch sollten die Ergebnisse aus dieser Arbeit bei der Verwendung von vergleichbaren Datensätzen reproduzierbar sein.

In den I-KNOW-Daten wurden ausschließlich Patienten mit erstmaligen, unilateralen Schlaganfällen im vorderen Stromgebiet eingeschlossen. Deshalb erlaubt dies keine Aussagen über die Prognosegüte der hier entwickelten Modelle in Bezug auf

Patienten mit sekundären, bilateralen und/oder Schlaganfällen im hinteren Stromgebiet, weil solche Daten nicht in die Modellierung eingeflossen sind.

Die tree-basierten Modelle in dieser Arbeit wurden nur für eine kleine Menge an Hyperparameterkombinationen im Rahmen einer Kreuzvalidierung validiert. Eine ausführliche Optimierung der Hyperparameter wurde nicht durchgeführt, weil das Ziel nicht in der maximalen Optimierung der Modelle bestand. Vielmehr galt es hier in erster Linie nachzuweisen, dass die räumlichen Faktoren zu einer konsistenten Verbesserung der Vorhersagen führen.

Während des Trainings wurde kein einheitliches, datengetriebenes Abbruchkriterium verwendet, um eine Überanpassung an die Trainingsdaten zu verhindern. Lediglich für XGBoost wurde die Anzahl der Entscheidungsbäume einheitlich auf zehn pro Modell begrenzt. Dies wirkt zwar einer Überanpassung an die Trainingsdaten entgegen, verschlechtert womöglich das Ergebnis der Validierungsmetriken, wenn dieser Hyperparameter restringiert wird.

Die in dieser Arbeit entwickelten Modelle wurden im Rahmen einer Kreuzvalidierung an Daten validiert, die an unterschiedlichen Kliniken erhoben wurden. Dennoch lohnt sich eine Überprüfung an einem unabhängigen Testdatensatz, um die externe Validität der Ergebnisse zu untermauern. Für eine Unterteilung in Trainings-, Validierungs- und Testdatensatz erweist sich der vorliegende Datensatz wahrscheinlich als zu klein, weshalb die Verallgemeinerbarkeit der entwickelten Machine-Learning-Modelle in Zukunft an einem unabhängigen Testdatensatz genauer zu untersuchen ist.

In dieser Arbeit wurden ausschließlich räumliche Faktoren einbezogen, die auf unterschiedlichen MNI-Koordinaten der Voxel oder der dort vorherrschenden Läsionswahrscheinlichkeit in den Trainingsdaten basieren. Atlasregionen oder der Gewebetyp der Voxel wurden insbesondere für die logistische Regression nicht als Feature validiert. Weiterhin wurden keine Voxelnachbarschaften modelliert.

In zukünftigen Studien sollte außerdem der finale Status der Rekanalisierung in die Modellierung einbezogen werden. Dadurch lässt sich die Prognose für einen unbehandelten Schlaganfall ohne einsetzende Rekanalisierung mit einer Prognose für ein Best-Case-Szenario gegenüberstellen, wenn z. B. durch eine Behandlung eine frühe Rekanalisierung durchgeführt wurde. Die Differenz zwischen beiden Prognosen entspricht dann einem Surrogat für die Penumbra, welches bei entsprechender Validität als Entscheidungsgrundlage für das weitere Therapievorgehen im Akutfall dient. Da

bei den Patienten der I-KNOW-Studie keine mechanische Rekanalisation erfolgte und damit keine angiographischen Informationen über den Rekanalisationszeitpunkt vorlagen, ließ sich diese Kontrastierung nicht durchführen.

Nach dem Vergleich der Ergebnisse der beiden Forschungsansätze werden diese nun im nächsten Abschnitt gemeinsam gegenüber den eingangs vorgestellten aktuellen Arbeiten kontrastiert, wobei hervorgehoben wird, welche Modellierungsschritte als Hautpursache für die unterschiedlichen Ergebnisse in Frage kommen. Um den Einfluss der jeweiligen Trainingsdatensätze besser einschätzen zu können, werden vor allem die Vergleiche mit den bereits etablierten Referenzmethoden aus den angeführten Arbeiten bemüht.

5.4 State of the Art

Rein auf Basis des Dice-Koeffizienten konnte der lokale Ansatz das beste Ergebnis der 15 Beteiligungen an den ISLES-Wettbewerben 2016 und 2017 übertreffen. Allerdings liegt schon der einfache globale Random Forest Benchmark aus dieser Studie mit einem Dice-Koeffizienten von 0,32 an der Spitze der ISLES Ergebnisse (2016: 0,31; 2017: 0,32). Deshalb muss berücksichtigt werden, dass die Infarkte im I-KNOW-Datensatz mit 0,58 ml im Durchschnitt um 20 ml größer sind als im ISLES-Datensatz und die Interrater-Reliabilität bei der Segmentierung der Ground Truth im Jahr 2016 lediglich einen Dice-Koeffizienten von 0,58 ergab.

Hier dienen daher die Dice-Ergebnisse der drei tree-basierten Einreichungen des Wettbewerbs als Referenz. Im Vergleich sind sie mit 0,30 bzw. 0,26 recht niedrig. Innerhalb des ISLES-Tableaus stehen sie allerdings im oberen Bereich. Während der lokale Random Forest keine signifikanten Verbesserungen bewirkte, zeigt der hybride Ansatz mit 0,35 einen deutlichen Anstieg. Gleiches gilt auch für die lokale (0,34) und hybride (0,35) logistische Regression. Diese relativen Verbesserungen durch die lokale und hybride Modellierung legen nahe, dass mit diesem Ansatz auch bei den ISLES-Wettbewerben einer der vorderen Plätze erzielt worden wäre.

In keines der ISLES-Prognose-Modelle flossen räumliche Features wie die in dieser Arbeit eingesetzten MNI-Koordinaten oder Wahrscheinlichkeitskarten ein.[30] Auch der XGBoost-Algorithmus wurde noch nicht verwendet. Das beste XGBoost-Modell aus

[30] Ähnliche Features wurden von [Robben et al., 2016] und [Mckinley et al., 2015] zur Segmentierung genutzt. Allerdings wurde der Einfluss der Features in diesen Arbeiten nicht diskutiert.

dieser Arbeit erreichte mit 0,40 einen deutlich höheren Dice-Koeffizienten als die Modellierungen der ISLES-Wettbewerbsbeiträge. Daher ist trotz der eingeschränkten Vergleichbarkeit von einem Steigerungspotential der ISLES-Ergebnisse durch die hier vorgestellte Modellierung auszugehen.

Ähnliches gilt auch für die Vergleiche zu den CNN in [Pinto et al., 2018a] und [Pinto et al., 2018b], bei denen die Zielmetrik um den Rekanalisationsstatus adjustiert wurde und die rohen PWI-Karten sowie einige Patientenfeatures einbezogen wurden. Während vor allem die rohen PWI-Karten ein vielversprechendes Feature für die tree-basierten Modelle sind, liegt in beiden Arbeiten das Validierungsergebnis für die ISLES-Daten bei einem Dice-Koeffizienten von 0,29. Damit ist das Ergebnis in Anbetracht der Unterschiede zu den I-KNOW-Daten etwa auf Augenhöhe mit dem lokalen und hybriden Ansatz, jedoch hinter der Modellierung von räumlichen Features in tree-basierten Modellen anzusiedeln.

Beim Vergleich mit dem von [Livne et al., 2018] trainierten XGBoost-Modell verhält es sich genau umgekehrt. Auf ihrem Datensatz erreichte schon die logistische Regression mit 0,87 eine deutlich höhere ROC AUC als in dieser Arbeit (0,81). Während die Größe der Infarkte nicht angegeben wurde, wurden jedoch über 25 Patientendatensätze wegen zu kleinen Läsionen aus ihrem Datensatz entfernt. Ebenfalls wurde die ROC AUC bei ihnen für das gesamte Hirnparenchym inkl. der kontralateralen Seite berechnet, was aufgrund einer gesteigerten True-Negative-Rate allgemein zu einem höheren Ergebnis führt als eine einseitige Evaluierung. Leider wurde kein Dice-Koeffizient angegeben, der diesbezüglich einen klareren Vergleich ermöglicht hätte. Trotz des signifikanten Ergebnisses ist die Verbesserung der ROC AUC für XGBoost gegenüber der logistischen Regression mit 0,01 eher gering.

Bei Verwendung derselben Features ist aufgrund der überaus deutlichen Verbesserungen der ROC AUC durch den Einbezug der räumlichen Features in die globale Modellierung (logistische Regression: +0,08; XGBoost: +0,06) von einem deutlichen Steigerungspotenzial auf ihrem Datensatz auszugehen. Neben den insgesamt 10 PWI-Karten bei [Livne et al., 2018] ist hier vor allem der Einbezug der akuten FLAIR-Sequenz zu nennen. Diese könnte auch für die globalen Modelle aus dieser Arbeit von Vorteil sein, da sie zusammen mit dem ADC das DWI-FLAIR-Mismatch modelliert, was einen Marker für das Alter von ADC-Läsionen darstellt und damit einen Hinweis auf die Überlebenschancen für das Gewebe bei einer Rekanalisation liefert.

Obwohl die Modelle in [Livne et al., 2018] auch gegenüber den zur Anreicherung der lokalen Modelle trainierten globalen Modelle bzgl. der ROC AUC (logistische Regression: 0,81; Random Forest: 0,79) überlegen sind, wurden in dieser Arbeit mithilfe der lokalen und hybriden Modellierung deutliche Verbesserungen um 0,05 bis 0,07 in dieser Metrik erzielt. Insbesondere mit den 195 Patientendatensätzen und den vermutlich etwas größeren Follow-up-Läsionen hätten deutlich mehr und robustere lokale Modelle trainiert werden können, für die mit der steigenden Patientenzahl auch weitere Features wie die FLAIR-Sequenz integriert werden könnten. Ebenso wie in dieser Arbeit scheint damit der lokale/hybride Ansatz einer tree-basierten Modellierung ohne räumliche Features überlegen zu sein.

Die von [Winder et al., 2019] publizierten Ergebnisse für K-Nearest-Neighbor-Modelle, logistische Regressionen und Random Forests, die unter verschiedenen Voraussetzungen trainiert wurden, liegen teils über den in dieser Arbeit erzielten Ergebnissen. Da die Läsionsgrößen mit durchschnittlich 53 ml nicht sehr stark von denen aus dieser Studie (58 ml) abweichen, bietet sich ein Vergleich an. Der Dice-Koeffizient ihrer Random Forests und logistischen Regressionen lag für vergleichbare Vorverarbeitungsschritte (Downsampling, Normalisierung an der kontralateralen Gehirnhälfte) mit durchschnittlich 0,47 bzw. 0,40 deutlich über denen der hybriden Modelle (je 0,35). Dies liegt hauptsächlich an der Berücksichtigung des Rekanalisationsstatus, der in anderen Arbeiten wie z. B. [Benzakoun et al., 2021] zu einem Anstieg des Dice-Koeffizienten um 0,07 geführt hat, sodass die hybride Modellierung trotz wesentlich weniger Features als bei [Winder et al., 2019] an die dort trainierte logistische Regression herankommt.

Ähnliches gilt für die tree-basierten Modelle aus dieser Arbeit, da ihnen ebenfalls der Rekanalisationsstatus fehlt. Hier ist davon auszugehen, dass die Verbesserung, die durch die räumlichen Features erzielt wurde leicht unter der Summe der Verbesserungen, die durch weitere Features bei [Winder et al., 2019] erzielt wurde, liegt. Diesbezüglich sind vor allem die Gehirnregion und die Gewebeart (graue vs. weiße Substanz) sowie der Abstand eines Voxels zum Infarktkern zu nennen.

Das von [A. Nielsen et al., 2018] auf 158 Patienten trainierte CNN (SegNet-Architektur) erreichte eine ROC AUC von 0,88±0,12. Sie verglichen ihr Modell u. a. mit einer logistischen Regression, deren ROC AUC mit 0,78±0,12 leicht unter der ROC AUC der globalen logistischen Regression liegt, die in dieser Arbeit zur Anreicherung des lokalen Modells verwendet wurde. Die ROC AUC der lokalen und hybriden

logistischen Regressionen liegen mit 0,86 und 0,87 nur knapp unter derjenigen des CNN. 105 der ursprünglich 222 verfügbaren Datensätze aus ihrer Arbeit stammen ebenfalls aus der I-KNOW-Studie. Interessant wäre deshalb der Vergleich mit den lokalen und hybriden Modellen unter Einbezug der weiteren Datensätze. Ebenfalls könnten die weiteren Biomarker aus ihrer Arbeit, wie z. B. die zerebrale Metabolisierungsrate von Sauerstoff ($CMRO_2$), in größeren lokalen Trainingsdatensätzen verwendet werden.

Die ROC AUC des besten XGBoost-Modells aus dieser Arbeit (ADC + PWI + MNI-Koordinaten) weist mit 0,89 denselben Wert wie das CNN auf. Im Gegensatz zu den fünf Tagen, die das Training des CNN beansprucht hat, benötigte das XGBoost-Modell lediglich 95 Sekunden auf einem gewöhnlichen Computer. Darin liegt das weitere Potenzial dieser Methode, da damit verschiedene Setups zeitökonomisch getestet werden können.

Das von [Yu et al., 2020] auf 182 Datensätzen trainierte CNN (U-Net-Architektur) ergab eine ähnlich hohe ROC AUC (0,89) und einen deutlich höheren Dice-Koeffizienten (0,53) als das beste Modell aus dieser Arbeit (0,40). Allerdings sind die Ergebnisse aufgrund des größeren Datensatzes von 182 Patienten und dessen Einfluss auf die Modellierung und den relativ großen Follow-up-Läsionen (Median FU-Läsionsvolumen von 54 ml) in ihrer Arbeit nicht direkt vergleichbar. Da in ihrer Arbeit lediglich DWI- und PWI-Features verwendet wurden, lässt sich in künftigen Studien ein Vergleich mit einem tree-basierten Modell mit MNI-Koordinaten auf demselben Datensatz relativ einfach und schnell durchführen. Dies sollte in zukünftigen Studien geschehen.

Der Datensatz bietet sich ebenfalls für Trainings des lokalen Ansatzes an, da er wesentlich mehr Patienten und fast doppelt so große Läsionen wie der in dieser Arbeit verwendete Trainingsdatensatz enthält, was das Training einer deutlich größeren Anzahl an lokalen Modellen ermöglichen würde.

Im erst kürzlich erprobten Ansatz von [Benzakoun et al., 2021] wurden sowohl Nachbarschafts- als auch kontralaterale Informationen basierend auf Patches in tree-basierte Modelle integriert und zusätzlich die MNI-Koordinaten inkludiert. Die Dice-Koeffizienten fallen mit 0,48 (XGBoost) und 0,47 (Random Forest) ebenfalls besser aus als die hier verwendeten lokalen und hybriden Modelle.

Ein Vorteil der Arbeit ist sicher die mit 394 Patienten sehr hohe Anzahl an verfügbaren Datensätzen. Ein Nachteil der Modellierung über die Integration der Features von Nachbarschaftsvoxeln ist, dass die Datensätze in [Benzakoun et al., 2021] aufgrund der hohen Featurezahl auf 5 % reduziert werden mussten. Deshalb wäre ein Downsampling vor allem von Nichtläsionsvoxeln sinnvoll, wie dies in dieser Arbeit und bei [Winder et al., 2019] geschehen ist. So ließen sich ausgeglichene Trainingsdatensätze erhalten. Da relativ kleine Patches verwendet wurden, die den Infarktkern in der Regel nicht beinhalten, ist es eine interessante Fragestellung, welchen zusätzlichen oder alternativen Nutzen der Abstand zum Infarktkern beisteuert, der bei [Winder et al., 2019] als Feature erprobte wurde.

6 Ausblick

Der Beitrag dieser Arbeit liegt in der Erprobung von zwei unterschiedlichen Ansätzen zur Integration von räumlichen Informationen in Modelle zur Vorhersage des Gewebeoutcomes beim ischämischen Schlaganfall (siehe Kapitel 3).

Im ersten Ansatz wurden die räumlichen Informationen über einen lokalen Modellierungsansatz integriert, bei dem ein Modell pro Gehirnposition trainiert wurde (siehe Abschnitt 3.4). Als Algorithmen kamen dabei die logistische Regression und Random Forest zum Einsatz. Dieser Ansatz wurde durch die Hinzunahme von globalen Prädiktionen nicht nur erweitert, sondern auch in eine hybride Modellierung umgewandelt. Insbesondere der hybride Ansatz führt zu einer deutlichen Verbesserung der Prognoseergebnisse gegenüber dem globalen Ansatz (siehe Abschnitt 4.2). Auf Basis dieses Ergebnisses empfiehlt sich in zukünftigen Studien mit größeren Patientenanzahlen die Nutzung der lokalen und hybriden Form der Modellierung, um dabei neben räumlichen Faktoren auch Patientenfeatures einzubeziehen, ohne diesbezüglich einen systematischen Bias zu erzeugen (siehe Abschnitt 5.1.7).

Beim zweiten Ansatz wurden räumliche Informationen als Features in globale Modelle integriert (siehe Abschnitt 3.5). Mit der Nutzung von MNI-Koordinaten sowie vorberechneten relativen Läsionshäufigkeiten pro Gehirnposition ergaben sich weitere signifikante Verbesserungen gegenüber einer lokalen und hybriden Modellierung (siehe Abschnitt 4.4). In allen globalen Modellen verhelfen die räumlichen Features zur Abgrenzung der gefährdeten Territorien des vorderen Stromgebiets vom restlichen Hirngewebe. Aufgrund der Eigenschaften von tree-basierten Modellen, nichtlineare Features zu verarbeiten und Interaktionseffekte automatisch zu erlernen, gelang diese Differenzierung wesentlich genauer als mittels logistischer Regression (siehe Abschnitte 4.3.1 und 5.2.2–5.2.3). Die besten Ergebnisse wurden insgesamt über den Einbezug von räumlichen Features mit dem XGBoost-Algorithmus für den globalen Ansatz erzielt (siehe Abschnitt 4.4).

In dieser Arbeit wurde nachvollziehbar, wie unterschiedlich räumliche Informationen in der logistischen Regression (siehe Abschnitt 5.2.2) versus in tree-basierten Modellen genutzt werden (siehe Abschnitt 5.2.3). Für XGBoost-Modelle wurden die Einflüsse der einzelnen Features sowohl auf Modell- als auch auf Voxel- und Patientenebene erfasst (siehe Abschnitt 4.3.2). Dabei wurde mit den Shapley-Werten erstmalig ein Framework zur Zerlegung von Vorhersagekarten in Beiträge einzelner Features erprobt, sodass die ähnlichen Effekte von MNI-Koordinaten und dem LP-Feature auf

tree-basierte Modelle visuell validiert werden konnten (siehe Abschnitt 4.3.2). Damit wurden die Anforderungen erfüllt, die sich aus den ISLES-Wettbewerben schlussfolgern ließen: Es galt, a priori bekannte physiologische Informationen zu Gehirnläsionen in die Modellierung miteinzubeziehen und dabei gleichzeitig die Transparenz und die Interpretierbarkeit der Modelle zu wahren.

Die Verbesserungen der Prognoseergebnisse durch den Einbezug von Läsionsverteilungen sind konsistent zu den Beobachtungen im Tierversuch bei [Shen und Duong, 2008] und in [Kemmling et al., 2015] beim Menschen. Um absolute State-of-the-Art-Ergebnisse zu erzielen, muss allerdings eine zusätzliche Modellierung der Voxelnachbarschaft wie in CNN oder bei [Benzakoun et al., 2021] vorgenommen werden, die MNI-Koordinaten mit einer patchbasierten Nachbarschaftsmodellierung in einem XGBoost-Modell kombinierten. Diese Kombination aus Orts- und Nachbarschaftsinformationen scheint vorteilhaft zu sein, sodass auch in CNN Läsionsverteilungen als Feature integriert werden sollten.

Da die in dieser Arbeit verwendeten I-KNOW-Daten Verschlüsse verschiedener Gefäße des vorderen Stromgebietes enthalten, fehlt den Modellen eine genaue Lokalisation des Verschlusses bzw. Informationen über das entsprechende Gefäß, die für eine individuellere Modellierung genutzt werden können. Zur genaueren Eingrenzung des Versorgungsgebietes sollten diese Informationen zusammen mit den Läsionshäufigkeitskarten der verschiedenen Verschlussorte in zukünftigen Studien integriert werden. Dann könnten durch Machine Learning die Territorien der größeren Gefäße in tree-basierten Modellen über die Interaktion aus Verschlussort und MNI-Koordinaten gelernt werden, was eine Weiterentwicklung des State of the Art ermöglichen würde. Eine automatische Bestimmung des Verschlussortes ist Gegenstand aktueller Forschung, wird jedoch insbesondere für Verschlüsse größerer Gefäße und deren Äste bereits erfolgreich eingesetzt [Dehkharghani et al., 2021].[31]

Innerhalb der Versorgungsgebiete spielt der zunächst unbekannte Rekanalisationsstatus eine entscheidende Rolle für die Entwicklung des gefährdeten Gewebes. In aktuellen Arbeiten werden deshalb separate Modelle für unterschiedliche Szenarien trainiert. Insbesondere bei einer gelungenen Rekanalisation kommt es auf den Zustand der Kollateralen und den Zeitpunkt der Rekanalisation an. Demnach sollte z. B. das

[31] Darüber hinaus können weitere Informationen über die Gefäße wie z. B. den Durchmesser oder den Verlauf von Arterien sowie die lokale Arteriendichte über einen zerebrovaskulären Atlas generiert werden (siehe z. B. [Forkert et al., 2013]).

bei [Galinovic et al., 2018] vorgeschlagene Verhältnis von CBF zu T_{max} als Marker für den Kollateralstatus in die Modellierung einfließen. Weiterhin gilt es, die FLAIR-Sequenz wie u. a. in [Livne et al., 2018] einzubeziehen, da sie in Kombination mit dem ADC eine Modellierung des DWI-FLAIR-Mismatches erlaubt, welches einen Marker für das Infarktalter bietet, und somit ggf. kritisches Gewebe identifiziert, das noch von einer frühen Rekanalisation profitieren würde.

Zur Bewertung des Einflusses einzelner (räumlicher) Features in zukünftigen Studien bieten sich die in dieser Arbeit verwendeten Shapley-Werte an. Mit dieser modellagnostischen Methode gelingt es, sowohl Aufschluss über die Einflüsse der Nachbarschafts- und kontralateralen Informationen in den tree-basierten Modellen bei [Benzakoun et al., 2021] zu erhalten als auch die Effekte der reinen Perfusionskarten in neuronalen Netzen bei [Pinto et al., 2018b] zu untersuchen. Bei der Integration weiterer Patientenfeatures (wie z. B. dem Rekanalisationsstatus bei [Benzakoun et al., 2021]) bieten Shapley-Werte eine Möglichkeit, deren Einfluss auf Voxelebene zu validieren, um so ggf. Anzeichen für einen systematischen Bias zu erkennen. Bei der Integration der Verschlusslokalisation in tree-basierte Modelle ist anzunehmen, dass die Shapley-Werte des Verschluss-Features und der MNI-Koordinaten das zugehörige Versorgungsgebiet anzeigen.

Für eine Modellierung von Patientenfeatures über den lokalen Ansatz sollte dieser in zukünftigen Studien weiter modifiziert werden. Hauptsächlich muss bei der Ergänzung ein globales Modell gewählt werden, das ebenfalls die räumliche Verteilung der Infarkte einbezieht, da der lokale Ansatz sonst vor allem außerhalb der Versorgungsgebiete vergleichsweise fehleranfällig wird. Im Idealfall wird allerdings über heterogenere Trainingsdaten das Training des lokalen Ansatzes auf dem gesamten MNI-Raum ermöglicht. Hier sollte auf Patientenebene der Ort des Gefäßverschlusses in die Modellierung integriert werden, da somit die Lage des Modells entscheidend an Relevanz gewinnt und vorhersagt, ob das Voxel zum Versorgungsgebiet des verschlossenen Gefäßes gehört oder nicht.

Bei der Entwicklung neuer Methoden sind zum Vergleich vor allem tree-basierte Modelle zu bevorzugen, die idealerweise räumliche sowie Nachbarschaftsinformationen wie bei [Benzakoun et al., 2021] enthalten. Ob sich tatsächlich – wie bislang angenommen – für die Vorhersage des Gewebeoutcomes eher Deep-Learning-Modelle eignen, da sie komplexe Features selbstständig aus Daten erlernen [Mouridsen et al., 2020], muss in Zukunft erst auf Basis von Studien mit größeren Patientenzahlen belegt

werden. Dennoch sind die Ergebnisse dieser Arbeit zusätzlich für Anwendungen wie Machine-Learning-gestützte in-silico-Trials von Bedeutung, da dort die Effektstärken neuer Therapiemethoden (z. B. Thrombektomie-Devices) auf Basis von kleineren Patientenkollektiven mithilfe von geeigneten Modellen wie Random Forests geschätzt werden können [Winder et al., 2021].

Für die Entwicklung verschiedener Methoden und den vereinheitlichten Vergleich auf einem einheitlichen Datensatz werden frei zugängliche Datensätze mit einer größeren Patientenanzahl als der ISLES-Datensatz benötigt. Damit könnte eine reproduzierbare Validierung anhand systematischer Benchmarks für verschiedene Algorithmen und Features stattfinden, die mittlerweile gut erforscht sind und als Vergleich für neue Methoden dienen, die auf diesem Datensatz entwickelt werden. Darüber hinaus sollten diese und andere Modelle systematisch zu Ensemble-Modellen kombiniert werden, sodass die unterschiedlichen Vorteile einzelner Algorithmen besser zur Geltung kommen. Die Daten sollten sowohl als Rohdaten als auch in vorprozessierter Form vorliegen, sodass es den Autoren neuer Studien selbst überlassen bleibt, ob sie sich auf Machine-Learning-Aspekte fokussieren oder zusätzliche Vorprozessierungsschritte anwenden. Weiterhin sollten Informationen zum Rekanalisationsstatus und dem Rekanalisationszeitpunkt im Datensatz enthalten sein, da diese sich maßgeblich auf den Krankheitsverlauf auswirken und somit bedeutende Implikationen für die Modellierung liefern.

Inwieweit die Ergebnisse aus dieser und ähnlichen Arbeiten Eingang in den klinischen Praxisalltag finden werden, hängt auch von rechtlichen Anforderungen ab, die derzeit diskutiert werden [Gerke et al., 2020]. Je mehr Transparenz bei der Modellierung vom Gesetzgeber gefordert wird, umso relevanter werden Interpretationstechniken wie z. B. die in dieser Arbeit vorgestellten Shapley-Werte und darauf basierende Karten zur Interpretation von Modellvorhersagen. Diese sollten nicht nur während der Entwicklung, sondern vor allem im Zuge der Etablierung von Modellen zusammen mit den bewährten DWI- und PWI-Karten zur manuellen Überprüfung automatisiert erzeugter Läsionsvorhersagen eingesetzt werden.

Der größte Hebel in der Akuttherapie des ischämischen Schlaganfalls liegt in der Rettung des bedrohten Gewebes durch die Thrombolyse und die Thrombektomie. Um die Abhängigkeit der engen Zeitfenster dieser Therapien in Zukunft weiter zu verringern und eine verbesserte personalisierte Therapieentscheidung zu treffen, müssen die hierzu benötigten aktuellen Verfahren der Gewebeprognose weiter optimiert wer-

den. Die in dieser Arbeit vorgestellten Methoden zur Integration von räumlichen Informationen in binäre Klassifikationsmodelle und die damit erzielten Verbesserungen tragen somit direkt dazu bei, dass zukünftig genauere Einschätzungen der Läsionsentwicklung für eine verbesserte Patientenauswahl getroffen werden können. Aufgrund der enormen Auswirkungen eines ischämischen Schlaganfalls auf die Gesundheit des Einzelnen und seiner extrem hohen Prävalenz in der Gesellschaft bedeuten selbst kleine Verbesserungen bei der Therapieentscheidung für eine große Patientenzahl eine erhebliche günstigere Rehabilitationsprognose und damit eine Entlastung des Gesundheitssystems.

Anhang

Tabelle A1: Hyperparametereinstellungen für XGBoost

Setting	eta	min_child_weight	subsample	colsample_bytree	max_depth	gamma	base_score
1	0,05	4	0,3	0,05	5	0,90	0,25
2	0,75	2	0,7	0,45	11	0,34	0,75
3	0,05	2	0,1	0,65	14	0,18	0,50
4	0,35	4	0,6	0,05	8	0,66	0,75
5	0,35	12	0,4	0,25	11	0,82	0,75
6	0,45	10	0,4	0,65	11	0,10	0,25
7	0,45	6	0,7	0,65	8	0,58	0,50

Abgesehen von den unterschiedlichen Hyperparametern der einzelnen Settings wurden für jedes Modell 10 Trees mit den folgenden Einstellungen trainiert: objective = binary: logistic; booster = gbtree. Eine Beschreibung der genauen Rolle jedes Parameters sowie seines möglichen Wertebereichs wird in der XGBoost-Dokumentation unter https://xgboost.readthedocs.io/en/latest/parameter.html angegeben (Quelle: Grosser et al., 2020a).

Tabelle A2: Strukturierte Hypothesentests für die Modelle aus dem lokalen und hybriden Forschungsansatz

Metrik	Ansatz Modell 1	Ansatz Modell 2	Algorithmus Modell 1	Algorithmus Modell 2	Effektstärke	P-Wert
ROC AUC	lokal	global	LR	LR	0,051398**	$3{,}027 \cdot 10^{-5}$
	lokal	hybrid			-0,019114**	0,003699
	global	hybrid			-0,623102**	$2{,}070 \cdot 10^{-10}$
	lokal	global	RF	RF	0,056749**	$3{,}308 \cdot 10^{-9}$
	lokal	hybrid			-0,013259**	$3{,}531 \cdot 10^{-7}$
	global	hybrid			-0,070000**	$1{,}318 \cdot 10^{-17}$
Dice-Koeffizient	lokal	global	LR	LR	0,015248	0,069941
	lokal	hybrid			-0,010785*	0,010774
	global	hybrid			-0,026033**	$2{,}904 \cdot 10^{-6}$
	lokal	global	RF	RF	-0,007841	0,310308
	lokal	hybrid			-0,041708**	$2{,}893 \cdot 10^{-14}$
	global	hybrid			-0,033867**	$2{,}950 \cdot 10^{-12}$
ROC AUC	lokal	lokal	LR	RF	0,015418**	$1{,}539 \cdot 10^{-5}$
	global	global			0,020768**	0,000145
	hybrid	hybrid			0,013071**	$2{,}054 \cdot 10^{-5}$
Dice-Koeffizient	lokal	lokal	LR	RF	0,026021**	$3{,}15710^{-8}$
	global	global			0,002933	0,576822
	hybrid	hybrid			-0,004902	0,304268
Sensitivität + Spezifität	lokal	global	LR	LR	0,058536**	0,000150
Sensitivität + Spezifität	lokal	hybrid			0,004254	0,532184
Sensitivität + Spezifität	global	hybrid			-0,054282**	$1{,}428 \cdot 10^{-8}$
Sensitivität + Spezifität	lokal	global	RF	RF	0,033964*	0,010920
Sensitivität + Spezifität	lokal	hybrid			-0,010275*	0,010275
Sensitivität + Spezifität	global	hybrid			-0,053138**	$9{,}664 \cdot 10^{-11}$
Sensitivität + Spezifität	lokal	lokal	LR	RF	0,043222**	$2{,}283 \cdot 10^{-9}$
Sensitivität + Spezifität	global	global			0,018650**	0,006793
Sensitivität + Spezifität	hybrid	hybrid			0,019794**	0,001025

Zweiseitige Hypothesentests für die sechs im Rahmen des lokalen Ansatzes trainierten Vorhersagemodelle. Für jede der Metriken wurden die Modelle je einmal nach Algorithmus (logistische Regression, Random Forest) unterteilt und innerhalb dieser Gruppen auf Unterschiede zwischen Modellen unterschiedlicher Ansätze (lokal, hybrid, global) untersucht und einmal nach Ansatz unterteilt und auf Unterschiede bzgl. des Algorithmus untersucht. Signifikante Unterschiede zwischen Modellen wurden mit einem Stern (*) bei einem Signifikanzniveau von $p < 0{,}05$ bzw. mit zwei Sternen (**) bei einem Signifikanzniveau von $p < 0{,}01$ gekennzeichnet. LR = logistische Regression, RF = Random Forest.

Tabelle A3: Medianwerte der einzelnen Koeffizienten der lokalen logistischen Regression pro Gehirnregion

VOI	Intercept	ADC	CBF	CBV	MTT	T_{max}
Nucleus caudatus	0,8011 ±1,4799	-0,0017 ±0,0013	-1,8685 ±1,3897	0,5217 ±0,6201	0,0415 ±0,0723	0,0217 ±0,0389
Zerebellum	-0,2827 ±1,6072	-0,002 ±0,0016	-0,2569 ±0,8302	-0,0837 ±0,4689	0,1319 ±0,106	0,0129 ±0,0413
Frontallappen	-0,87 ±1,3517	-0,001 ±0,0014	-0,5195 ±1,194	-0,184 ±0,641	0,1818 ±0,1313	0,0583 ±0,0516
Insula	-0,7254 ±1,3496	-0,0009 ±0,001	-0,3134 ±0,7195	0,0431 ±0,3216	0,0766 ±0,0506	0,0551 ±0,0276
Okzipitallappen	0,2258 ±1,606	-0,0022 ±0,0016	-0,6754 ±1,6441	0,0874 ±0,7144	0,1407 ±0,1163	0,0532 ±0,0449
Parietallappen	-0,3865 ±1,2535	-0,0012 ±0,0012	-0,3958 ±1,1542	-0,1272 ±0,705	0,1205 ±0,1022	0,0489 ±0,0354
Putamen	0,3894 ±1,3084	-0,0018 ±0,0012	-0,8067 ±0,8587	0,1076 ±0,3898	0,0775 ±0,074	0,044 ±0,0352
Temporallappen	-0,4691 ±1,0631	-0,0011 ±0,0011	-0,4013 ±0,6622	0,002 ±0,3294	0,1315 ±0,078	0,0452 ±0,0287
Thalamus	1,195 ±1,283	-0,0022 ±0,0014	-0,528 ±2,0799	-0,2935 ±0,8333	0,1895 ±0,114	0,0298 ±0,0482
Global	-0,043 ±0,027	-0,0031 ±0	-0,1184 ±0,0117	-0,2889 ±0,0211	0,2236 ±0,0021	0,0856 ±0,0001

Medianwerte der Koeffizienten der lokalen logistischen Regressionsmodelle, die an Positionen innerhalb der jeweiligen Gehirnregion trainiert wurden. Da die Werte der Koeffizienten keiner Normalverteilung folgten, wurden für die lokalen Modellkoeffizienten Medianwerte und die mittlere absolute Abweichung vom Median angegeben. Zum Vergleich wurden außerdem die eindeutigen Punktschätzer der Koeffizienten und deren Standardabweichungen für die globale logistische Regression in der letzten Zeile der Tabelle angegeben (Quelle: Grosser et al., 2020b).

Tabelle A4: Durchschnittliche ROC AUCs und Dice-Koeffizienten für die globalen Modelle nach Integration räumlicher Features

Modell	Features	Setting	Ø ROC AUC (Training)	Ø ROC AUC	Ø Dice-Koeffizient	Trainingszeit (s)
LR	ADC + PWI		0,817±0,001	0,813±0,107**	0,317±0,220**	41
LR	ADC + PWI + MNI		0,849±0,001	0,827±0,100**	0,292±0,229**	48
LR	ADC + PWI + LP		0,901±0,001	0,874±0,108**	0,319±0,238**	45
LR	ADC + PWI + MNI + LP		0,902±0,001	0,877±0,099**	0,322±0,232**	44
RF	ADC + PWI		0,887±0,001	0,826±0,104**	0,341±0,218**	614
RF	ADC + PWI + MNI		0,969±0,000	0,891±0,092	0,383±0,226**	628
RF	ADC + PWI + LP		0,950±0,001	0,883±0,104**	0,371±0,227**	622
RF	ADC + PWI + MNI + LP		0,980±0,000	0,889±0,092	0,368±0,228**	758
XGB	ADC + PWI	1	0,819±0,010	0,814±0,118**	0,321±0,222**	45
XGB	ADC + PWI	2	0,844±0,003	0,826±0,104**	0,337±0,218**	98
XGB	ADC + PWI	3	0,843±0,002	0,827±0,108**	0,337±0,222**	146
XGB	ADC + PWI	4	0,827±0,008	0,825±0,107**	0,325±0,223**	62
XGB	ADC + PWI	5	0,827±0,009	0,821±0,113**	0,322±0,224**	80
XGB	ADC + PWI	6	0,849±0,001	0,828±0,105**	0,346±0,221**	107
XGB	ADC + PWI	7	0,846±0,001	0,830±0,105**	0,346±0,220**	77
XGB	ADC + PWI + MNI	1	0,859±0,017	0,852±0,106**	0,336±0,214**	44
XGB	ADC + PWI + MNI	2	0,923±0,004	0,883±0,092**	0,374±0,221**	103
XGB	ADC + PWI + MNI	3	0,926±0,004	**0,893±0,085**	0,387±0,213	179
XGB	ADC + PWI + MNI	4	0,869±0,015	0,850±0,113**	0,329±0,224**	59
XGB	ADC + PWI + MNI	5	0,897±0,009	0,880±0,091**	0,355±0,216**	93
XGB	ADC + PWI + MNI	6	0,934±0,002	0,889±0,093*	0,387±0,222	122
XGB	ADC + PWI + MNI	7	0,922±0,002	0,890±0,092	0,386±0,218	94
XGB	ADC + PWI + LP	1	0,875±0,029	0,867±0,100**	0,353±0,224**	55
XGB	ADC + PWI + LP	2	0,917±0,004	0,884±0,105*	0,379±0,228**	121
XGB	ADC + PWI + LP	3	0,909±0,007	0,890±0,093	0,384±0,225*	175
XGB	ADC + PWI + LP	4	0,893±0,021	0,878±0,102**	0,367±0,220**	78
XGB	ADC + PWI + LP	5	0,888±0,029	0,875±0,102**	0,365±0,222**	93
XGB	ADC + PWI + LP	6	0,920±0,002	0,886±0,101*	0,387±0,224*	122
XGB	ADC + PWI + LP	7	0,918±0,001	0,888±0,101	**0,395±0,229**	95
XGB	ADC + PWI + MNI + LP	1	0,880±0,018	0,869±0,094**	0,336±0,225**	55
XGB	ADC + PWI + MNI + LP	2	0,936±0,002	0,881±0,097**	0,374±0,220**	144
XGB	ADC + PWI + MNI + LP	3	0,932±0,001	0,888±0,104	0,379±0,226**	202
XGB	ADC + PWI + MNI + LP	4	0,889±0,017	0,876±0,094**	0,351±0,226**	80
XGB	ADC + PWI + MNI + LP	5	0,910±0,006	0,887±0,097*	0,375±0,231**	120
XGB	ADC + PWI + MNI + LP	6	0,937±0,001	0,887±0,098*	0,386±0,224	157
XGB	ADC + PWI + MNI + LP	7	0,926±0,001	0,890±0,097	0,381±0,228**	114

Die besten Ergebnisse in Bezug auf die einzelnen Metriken sind jeweils fett gedruckt hervorgehoben. Signifikante Unterschiede einzelner Modelle zum jeweils besten Modell bzgl. der jeweiligen Metrik wurden mit einem einseitigen T-Test berechnet und mit einem Stern (*) bei einem Signifikanzniveau von $p < 0,05$ bzw. mit zwei Sternen (**) bei einem Signifikanzniveau von $p < 0,01$ gekennzeichnet. Nominale p-Werte wurden ohne Korrektur für multiples Testen berechnet, ähnlich wie in [Maier et al., 2015]. LR = logistische Regression, RF = Random Forest, XGB = XGBoost, ADC = apparent diffusion coefficient, PWI = perfusion-weighted MRI parameters, MNI = Montreal Neurological Institute atlas coordinates, LP = lesion probability (Quelle: Grosser et al., 2020a).

Tabelle A5: Strukturierte Hypothesentests für die Modelle aus dem globalen Ansatz mit räumlichen Features

Metrik	Features Modell 1	Features Modell 2	Algorithmus Modell 1	Algorithmus Modell 2	Effektstärke	P-Wert
ROC AUC	ADC + PWI	ADC + PWI + MNI	LR	LR	-0,013260	0,119081
	ADC + PWI	ADC + PWI + LP			-0,060944**	$2{,}648 \cdot 10^{-6}$
	ADC + PWI	ADC + PWI + MNI + LP			-0,063926**	$5{,}120 \cdot 10^{-7}$
	ADC + PWI + MNI	ADC + PWI + LP			-0,047685**	0,000140
	ADC + PWI + MNI	ADC + PWI + LP + MNI			-0,050666**	$5{,}073 \cdot 10^{-6}$
	ADC + PWI + LP	ADC + PWI + LP + MNI			-0,002982	0,226387
Dice-Koeffizient	ADC + PWI	ADC + PWI + MNI			0,025143**	0,004267
	ADC + PWI	ADC + PWI + LP			-0,001256	0,931697
	ADC + PWI	ADC + PWI + MNI + LP			-0,004396	0,748905
	ADC + PWI + MNI	ADC + PWI + LP			-0,026399	0,158526
	ADC + PWI + MNI	ADC + PWI + LP + MNI			-0,029539	0,090192
	ADC + PWI + LP	ADC + PWI + LP + MNI			-0,003140	0,198123
ROC AUC	ADC + PWI	ADC + PWI + MNI	RF	RF	-0,064725**	$9{,}120 \cdot 10^{-9}$
	ADC + PWI	ADC + PWI + LP			-0,056369**	$3{,}360 \cdot 10^{-6}$
	ADC + PWI	ADC + PWI + MNI + LP			-0,062155**	$5{,}532 \cdot 10^{-8}$
	ADC + PWI + MNI	ADC + PWI + LP			0,008356**	0,003481
	ADC + PWI + MNI	ADC + PWI + LP + MNI			0,002569*	0,049724
	ADC + PWI + LP	ADC + PWI + LP + MNI			-0,005787*	0,039838
Dice-Koeffizient	ADC + PWI	ADC + PWI + MNI			-0,041540**	$9{,}622 \cdot 10^{-6}$
	ADC + PWI	ADC + PWI + LP			-0,029489**	0,004123
	ADC + PWI	ADC + PWI + MNI + LP			-0,027279**	0,014594
	ADC + PWI + MNI	ADC + PWI + LP			0,012051**	0,000967
	ADC + PWI + MNI	ADC + PWI + LP + MNI			0,014262**	0,000662
	ADC + PWI + LP	ADC + PWI + LP + MNI			0,002210	0,527909
ROC AUC	ADC + PWI	ADC + PWI + MNI	XGBoost	XGBoost	-0,063468**	$1{,}012 \cdot 10^{-10}$
	ADC + PWI	ADC + PWI + LP			-0,058029**	$9{,}276 \cdot 10^{-7}$
	ADC + PWI	ADC + PWI + MNI + LP			-0,057323**	$1{,}064 \cdot 10^{-6}$
	ADC + PWI + MNI	ADC + PWI + LP			0,005439	0,142616
	ADC + PWI + MNI	ADC + PWI + LP + MNI			0,006145	0,059730
	ADC + PWI + LP	ADC + PWI + LP + MNI			0,000706	0,671192
Dice-Koeffizient	ADC + PWI	ADC + PWI + MNI			-0,040450**	$6{,}867 \cdot 10^{-8}$
	ADC + PWI	ADC + PWI + LP			-0,048108**	$4{,}999 \cdot 10^{-8}$
	ADC + PWI	ADC + PWI + MNI + LP			-0,039788**	$7{,}141 \cdot 10^{-5}$
	ADC + PWI + MNI	ADC + PWI + LP			-0,007659	0,135520
	ADC + PWI + MNI	ADC + PWI + LP + MNI			0,000662	0,912299
	ADC + PWI + LP	ADC + PWI + LP + MNI			0,008321	0,197718
ROC AUC	ADC + PWI	ADC + PWI	LR	RF	-0,012922**	0,000400
	ADC + PWI + MNI	ADC + PWI + MNI			-0,064387**	$1{,}332 \cdot 10^{-8}$
	ADC + PWI + LP	ADC + PWI + LP			-0,008346**	0,000887
	ADC + PWI + MNI + LP	ADC + PWI + MNI + LP			-0,011151**	0,000339
	ADC + PWI	ADC + PWI	LR	XGBoost	-0,016228**	$9{,}591 \cdot 10^{-6}$
	ADC + PWI + MNI	ADC + PWI + MNI			-0,066436**	$4{,}409 \cdot 10^{-11}$
	ADC + PWI + LP	ADC + PWI + LP			-0,013312**	$5{,}165 \cdot 10^{-9}$
	ADC + PWI + MNI + LP	ADC + PWI + MNI + LP			-0,009625**	0,000753
	ADC + PWI	ADC + PWI	RF	XGBoost	-0,003306**	0,000392
	ADC + PWI + MNI	ADC + PWI + MNI			-0,002050	0,420222
	ADC + PWI + LP	ADC + PWI + LP			-0,004967**	$1{,}619 \cdot 10^{-5}$
	ADC + PWI + MNI + LP	ADC + PWI + MNI + LP			0,001526	0,360903
Dice-Koeffizient	ADC + PWI	ADC + PWI	LR	RF	-0,023888**	0,000169
	ADC + PWI + MNI	ADC + PWI + MNI			-0,090571**	$3{,}023 \cdot 10^{-9}$
	ADC + PWI + LP	ADC + PWI + LP			-0,052120**	$2{,}188 \cdot 10^{-8}$
	ADC + PWI + MNI + LP	ADC + PWI + MNI + LP			-0,046770**	$2{,}308 \cdot 10^{-8}$

Metrik	Features Modell 1	Features Modell 2	Algorithmus Modell 1	Algorithmus Modell 2	Effektstärke	P-Wert
ROC AUC	ADC + PWI	ADC + PWI	LR	XGBoost	-0,029132**	$2{,}156 \cdot 10^{-5}$
	ADC + PWI + MNI	ADC + PWI + MNI			-0,094724**	$9{,}176 \cdot 10^{-11}$
	ADC + PWI + LP	ADC + PWI + LP			-0,075984**	$5{,}116 \cdot 10^{-10}$
	ADC + PWI + MNI + LP	ADC + PWI + MNI + LP			-0,064523**	$2{,}744 \cdot 10^{-12}$
Dice-Koeffizient	ADC + PWI	ADC + PWI	RF	XGBoost	-0,005244	0,108279
	ADC + PWI + MNI	ADC + PWI + MNI			-0,004153	0,380457
	ADC + PWI + LP	ADC + PWI + LP			-0,023863**	$4{,}016 \cdot 10^{-5}$
	ADC + PWI + MNI + LP	ADC + PWI + MNI + LP			-0,017753**	$7{,}031 \cdot 10^{-5}$

Zweiseitige gepaarte Hypothesentests für die vier logistischen Regressionsmodelle, die vier Random Forests und die vier (pro Featurekombination) besten XGBoost-Modelle, die im Rahmen des globalen Ansatzes mit räumlichen Features trainiert wurden. Für jede der Metriken wurden die Modelle je einmal nach Algorithmus (logistische Regression, Random Forest, XGBoost) unterteilt und innerhalb dieser Gruppen auf Unterschiede zwischen den verschiedenen Featurekombinationen untersucht und einmal nach Featurekombination unterteilt und auf Unterschiede bzgl. des Algorithmus untersucht. Signifikante Unterschiede zwischen Modellen wurden mit einem Stern (*) bei einem Signifikanzniveau von $p < 0{,}05$ bzw. mit zwei Sternen (**) bei einem Signifikanzniveau von $p < 0{,}01$ gekennzeichnet. LR = logistische Regression, RF = Random Forest.

Tabelle A6: Strukturierte Hypothesentests für den Vergleich der Modelle aus dem lokalen bzw. hybriden Ansatz und dem globalen Ansatz mit räumlichen Features

Metrik	Algorithmus	Ansatz Modell 1	Features globaler Ansatz	Effektstärke	P-Wert
ROC AUC	LR	global	DWI + PWI	-0,004022	0,790829
	RF			-0,037712**	0,009970
Dice-Koeffizient	LR			0,004962	0,879611
	RF			-0,021858	0,494468
ROC AUC	LR	lokal	DWI + PWI + MNI	0,034117**	0,009515
			DWI + PWI + LP	-0,013568	0,368109
			DWI + PWI + MNI + LP	-0,016550	0,239074
Dice-Koeffizient	LR	lokal	DWI + PWI + MNI	0,045353	0,164967
			DWI + PWI + LP	0,018954	0,575436
			DWI + PWI + MNI + LP	0,015814	0,635687
ROC AUC	LR	hybrid	DWI + PWI + MNI	0,045028**	0,000287
			DWI + PWI + LP	-0,002657	0,849452
			DWI + PWI + MNI + LP	-0,005638	0,664139
Dice-Koeffizient	LR	hybrid	DWI + PWI + MNI	0,056137	0,091716
			DWI + PWI + LP	0,029738	0,38589
			DWI + PWI + MNI + LP	0,026599	0,431453
ROC AUC	RF	lokal	DWI + PWI + MNI	-0,045688**	0,000834
			DWI + PWI + LP	-0,037332**	0,008129
			DWI + PWI + MNI + LP	-0,043119**	0,001675
Dice-Koeffizient	RF	lokal	DWI + PWI + MNI	-0,071239*	0,025218
			DWI + PWI + LP	-0,059188	0,063501
			DWI + PWI + MNI + LP	-0,056978	0,075652
ROC AUC	RF	hybrid	DWI + PWI + MNI	-0,032429*	0,011041
			DWI + PWI + LP	-0,024073	0,070032
			DWI + PWI + MNI + LP	-0,029860*	0,019835
Dice-Koeffizient	RF	hybrid	DWI + PWI + MNI	-0,029531	0,369316
			DWI + PWI + LP	-0,017480	0,596216
			DWI + PWI + MNI + LP	-0,015270	0,645692

Zweiseitige Hypothesentests zum Vergleich von Modellen mit und ohne räumlichen Informationen bzgl. des Dice-Koeffizienten und der ROC AUC. Es wurden ausschließlich Modelle, die mittels desselben Algorithmus trainiert wurden, verglichen. Vier Tests wurden durchgeführt, um globale Modelle zwischen den beiden Ansätzen zu vergleichen (einer pro Algorithmus und Metrik). Für Modelle, die räumliche Informationen beinhalten, wurden 24 Tests durchgeführt. Hierzu wurde jedes der lokalen und hybriden Modelle mit den entsprechenden drei globalen Modellen mit räumlichen Informationen (einem pro Featurekombination) verglichen. Signifikante Unterschiede zwischen Modellen wurden mit einem Stern (*) bei einem Signifikanzniveau von $p < 0,05$ bzw. mit zwei Sternen (**) bei einem Signifikanzniveau von $p < 0,01$ gekennzeichnet. LR = logistische Regression, RF = Random Forest.

Literaturverzeichnis

Aho, K., Harmsen, P., Hatano, S., Marquardsen, J., Smirnov, V. und Strasser, T. (1980). **Cerebrovascular disease in the community: Results of a WHO collaborative study.** *Bulletin of the World Health Organization*, 58(1):113–30.

Alawneh, J., Jones, P., Mikkelsen, I., Cho, T., Siemonsen, S., Mouridsen, K., Ribe, L., Morris, R., Hjort, N., Antoun, N., Gillard, J., Fiehler, J., Nighoghossian, N., Warburton, E., Ostergaard, L. und Baron, J. (2011). **Infarction of "non-core-non-penumbral" tissue after stroke: Multivariate modelling of clinical impact**. *Brain*, 134(6):1765–76.

Albers, G., Caplan, L., Easton, J., Fayad, P., Mohr, J., Saver, J. und Sherman, D. (2002). **Transient ischemic attack – proposal for a new definition**. *New England Journal of Medicine*, 347(21):1713–16.

Andersen, K., Olsen, T., Dehlendorff, C. und Kammersgaard, L. (2009). **Hemorrhagic and ischemic strokes compared: Stroke severity, mortality, and risk factors**. *Stroke*, 40(6):2068–72.

Andjelkovic, A., Stamatovic, S., Phillips, C., Martinez-Revollar, G. und Keep, R. (2020). **Modeling blood-brain barrier pathology in cerebrovascular disease in vitro: current and future paradigms**. *Fluids and Barriers of the CNS*, 17(1):44.

Appelros, P., Stegmayr, B. und Terent, A. (2009). **Sex differences in stroke epidemiology: A systematic review**. *Stroke*, 40(4):1082–90.

Arakawa, S., Wright, P., Koga, M., Phan, T., Reutens, D., Lim, I., Gunawan, M., Ma, H., Perera, N., Ly, J., Zavala, J., Fitt, G. und Donnan, G. (2006). **Ischemic thresholds for gray and white matter: A diffusion and perfusion magnetic resonance study**. *Stroke*, 37(5):1211–16.

Avants, B., Kandel, B., Duda, J., Cook, P., Tustison, N. und Shrinidhi, K. (2017). **ANTsR: ANTs in R: quantification tools for biomedical images**. https://github.com/antsx/antsr.

Badrinarayanan, V., Kendall, A. und Cipolla, R. (2017). **SegNet: A deep convolutional encoder-decoder architecture for image segmentation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–95.

Baron, J. (2018). **Protecting the ischaemic penumbra as an adjunct to thrombectomy for acute stroke**. *Nature Reviews Neurology*, 14(6):325–37.

Beleites, C. (2016). **arrayhelpers: Convenience Functions for Arrays**. https://cran.r-project.org/package=arrayhelpers.

Benzakoun, J., Charron, S., Turc, G., Hassen, W., Legrand, L., Boulouis, G., Naggara, O., Baron, J., Thirion, B. und Oppenheim, C. (2021). **Tissue outcome prediction in hyperacute ischemic stroke: Comparison of machine learning models**. *Journal of Cerebral Blood Flow and Metabolism*, 3085–96.

Berkefeld, J. und Neumann-Haefelin, T. (2009). **Diagnostik der zerebralen Ischämie: Wann CT, wann MRT?** *Der Radiologe*, 49(4):299–304.

Berner, L.-P., Cho, T.-H., Haesebaert, J., Bouvier, J., Wiart, M., Hjort, N., Mikkelsen, I., Derex, L., Thomalla, G., Pedraza, S., Østergaard, L., Baron, J.-C., Nighoghossian, N. und Berthezène, Y. (2016). **MRI assessment of ischemic lesion evolution within white and gray matter**. *Cerebrovascular Diseases*, 41(5–6):291–97.

Billebaut, B. und Wameling, J. (2012). **Pulssequenzen und Kontrast – Magnetresonanztomografie Teil III**. *Radiopraxis*, 5(1):11–22.

Boehme, A., Esenwa, C. und Elkind, M. (2017). **Stroke risk factors, genetics, and prevention**. *Circulation Research*, 120(3):472–95.

Breiman, L. (2001). **Random Forests**. *Machine Learning*, 45(1):5–32.

Brown, M. und Semelka, R. (2003). **MRI**. John Wiley & Sons, Inc.

Bundesministerium für Bildung und Forschung. (2012). **Der Schlaganfall: Forschung – Diagnose – Therapie**. https://www.schlaganfallcentrum.de/fileadmin/redaktion/csb/pdf_flyerundbroschueren/bmbf_schlaganfallbroschuere.pdf.

Busch, M. und Kuhnert, R. (2017). **12-Month prevalence of stroke or chronic consequences of stroke in Germany**. *Journal of Health Monitoring*, 2(1):64–69.

Campbell, B., Christensen, S., Tress, B., Churilov, L., Desmond, P., Parsons, M., Barber, P., Levi, C., Bladin, C., Donnan, G. und Davis, S. (2013). **Failure of collateral blood flow is associated with infarct growth in ischemic stroke**. *Journal of Cerebral Blood Flow & Metabolism*, 33(8):1168–72.

Campbell, B. und Parsons, M. (2018). **Imaging selection for acute stroke intervention**. *International Journal of Stroke*, 13(6):554–67.

Catanese, L., Tarsia, J. und Fisher, M. (2017). **Acute ischemic stroke therapy overview**. *Circulation Research*, 120(3):541–58.

Chen, C., Bivard, A., Lin, L., Levi, C., Spratt, N. und Parsons, M. (2017). **Thresholds for infarction vary between gray matter and white matter in acute ischemic stroke: A CT perfusion study**. *Journal of Cerebral Blood Flow & Metabolism*, 39(3):536–46.

Chen, T. und Guestrin, C. (2016). **XGBoost: A scalable tree boosting system**. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94.

Chen, T., He, T., Benesty, M., Khotilovich, V. und Tang, Y. (2017). **xgboost: Extreme gradient boosting**. https://cran.r-project.org/package=xgboost.

Cheng, B. (2021). **Akuttherapie des Schlaganfalls mit unbekanntem Zeitfenster**. *InFo Neurologie + Psychiatrie*, 23(6):34–45.

Cheng, B., Forkert, N., Zavaglia, M., Hilgetag, C., Golsari, A., Siemonsen, S., Fiehler, J., Pedraza, S., Puig, J., Cho, T.-H., Alawneh, J., Baron, J.-C., Ostergaard, L., Gerloff, C. und Thomalla, G. (2014). **Influence of stroke infarct location on functional outcome measured by the modified Rankin Scale**. *Stroke*, 45(6):1695–1702.

Chollet, F. und Allaire, J. (2018). **Deep Learning with R**. Birmingham: Manning Publications.

Cossey, T. und Gonzales, N. (2015). **Thrombolysis-related hemorrhage: Can we make thrombolysis safer?** *JAMA Neurology*, 72(12):1416–18.

Davis, S. und Donnan, G. (2014). **Time is penumbra: Imaging, selection and outcome**. *Cerebrovascular Diseases*, 38(1):59–72.

Dehkharghani, S., Lansberg, M., Venkatsubramanian, C., Cereda, C., Lima, F., Coelho, H., Rocha, F., Qureshi, A., Haerian, H., Mont'Alverne, F., Copeland, K. und Heit, J. (2021). **High-performance automated anterior circulation CT angiographic clot detection in acute stroke: A multireader comparison**. *Radiology*, 298(3):665–70.

Dombrowski, S., White, M., Mackintosh, J., Gellert, P., Araujo-Soares, V., Thomson, R., Rodgers, H., Ford, G. und Sniehotta, F. (2014). **The stroke "Act FAST" campaign: remembered but not understood?** *International Journal of Stroke*, 10(3):324–30.

Dowle, M. und Srinivasan, A. (2020). **data.table: Extension of data.frame**. https://github.com/rdatatable/data.table.

Easton, J., Saver, J., Albers, G., Alberts, M., Chaturvedi, S., Feldmann, E., Hatsukami, T., Higashida, R., Johnston, S., Kidwell, C., Lutsep, H., Miller, E. und Sacco, R. (2009). **Definition and evaluation of transient ischemic attack: a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardio**. *Stroke*, 40(6):2276–93.

Ermine, C., Bivard, A., Parsons, M. und Baron, J.-C. (2020). **The ischemic penumbra: From concept to reality**. *International Journal of Stroke*, 16(5):497–509.

Evans, M., White, P., Cowley, P. und Werring, D. (2017). **Revolution in acute ischaemic stroke care: A practical guide to mechanical thrombectomy**. *Practical Neurology*, 17(4):252–65.

Faiss, J., Busse, O. und Ringelstein, E. (2008). **Aufgaben und Ausstattung einer Stroke-Unit: Weiterentwicklung des Stroke-Unit-Konzeptes in Deutschland**. *Nervenarzt*, 79(4):480–82.

Feigin, V., Norrving, B., Sudlow, C. und Sacco, R. (2018). **Updated criteria for population-based stroke and transient ischemic attack incidence studies for the 21st century**. *Stroke*, 49(9):2248–55.

Flottmann, F., Broocks, G., Faizy, T., Ernst, M., Forkert, N., Grosser, M., Thomalla, G., Siemonsen, S., Fiehler, J. und Kemmling, A. (2017). **CT-perfusion stroke imaging: A threshold free probabilistic approach to predict infarct volume compared to traditional ischemic thresholds**. *Scientific Reports*, 7(1):6679.

Flottmann, F., Leischner, H., Broocks, G., Nawabi, J., Bernhardt, M., Faizy, T., Deb-Chatterji, M., Thomalla, G., Fiehler, J. und Brekenfeld, C. (2018). **Recanalization rate per retrieval attempt in mechanical thrombectomy for acute ischemic stroke**. *Stroke*, 49(10):2523–25.

Forkert, N., Cheng, B., Kemmling, A., Thomalla, G. und Fiehler, J. (2014). **ANTONIA perfusion and stroke: A software tool for the multi-purpose analysis of MR perfusion-weighted datasets and quantitative ischemic stroke assessment**. *Methods of Information in Medicine*, 53(6):469–81.

Forkert, N., Fiehler, J., Suniaga, S., Wersching, H., Knecht, S. und Kemmling, A. (2013). **A statistical cerebroarterial atlas derived from 700 MRA datasets**. *Methods of Information in Medicine*, 467–74.

Forkert, N., Säring, D., Fiehler, J., Illies, T., Möller, D. und Handels, H. (2009). **Hämodynamische Analyse und Klassifikation der Gefäßstrukturen bei Patienten mit zerebralen arteriovenösen Malformationen**. *GMS Medizinische Informatik, Biometrie und Epidemiologie*, 5(3):Doc19.

Friedrich, B., Boeckh-Behrens, T., Krüssmann, V., Mönch, S., Kirschke, J., Kreiser, K., Berndt, M., Lehm, M., Wunderlich, S., Zimmer, C., Kaesmacher, J. und Maegerlein, C. (2020). **A short history of thrombectomy – Procedure and success analysis of different endovascular stroke treatment techniques**. *Interventional Neuroradiology*, 27(2):249–56.

Galinovic, I., Kochova, E., Khalil, A., Villringer, K., Piper, S. und Fiebach, J.B. (2018). **The ratio between cerebral blood flow and Tmax predicts the quality of collaterals in acute ischemic stroke**. *PLoS ONE*, 13(1):e0190811.

Gerke, S., Minssen, T. und Cohen, G. (2020). **Ethical and legal challenges of artificial intelligence-driven healthcare**. In *Artificial Intelligence in Healthcare*, 295–336. Elsevier.

Goodfellow, I., Bengio, Y. und Courville, A. (2016). **Deep learning**. MIT Press.

Goyal, M., Ospel, J., Menon, B., Almekhlafi, M., Jayaraman, M., Fiehler, J., Psychogios, M., Chapot, R., Lugt, A. van der, Liu, J., Yang, P., Agid, R., Hacke, W., Walker, M., Fischer, U., Asdaghi, N., McTaggart, R., … Fisher, M. (2020). **Challenging the ischemic core concept in acute ischemic stroke imaging**. *Stroke*, 51(10):3147–55.

Grosser, M., Gellißen, S., Borchert, P., Sedlacik, J., Nawabi, J., Fiehler, J. und Forkert, N. (2020a). **Improved multi-parametric prediction of tissue outcome in acute ischemic stroke patients using spatial features**. *PLoS ONE*, 15(1):e0230653.

Grosser, M., Gellißen, S., Borchert, P., Sedlacik, J., Nawabi, J., Fiehler, J. und Forkert, N.D. (2020b). **Localized prediction of tissue outcome in acute ischemic stroke patients using diffusion- and perfusion-weighted MRI datasets**. *PLoS ONE*, 15(11):e0241917.

Hand, D. und Vinciotti, V. (2003). **Local versus global models for classification problems**. *The American Statistician*, 57(2):124–31.

Hankey, G. (2017). **Stroke**. *The Lancet*, 389(10069):641–54.

Hastie, T., Tibshirani, R. und Friedman, J. (2009). **The elements of statistical learning: data mining, inference and prediction**. 2. Aufl. Springer.

Heit, J. und Wintermark, M. (2016). **Perfusion computed tomography for the evaluation of acute ischemic stroke strengths and pitfalls**. *Stroke*, 47(4):1153–58.

Higashida, R. und Furlan, A. (2003). **Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke**. *Stroke*, 34(8).

Hosmer, D., Lemeshow, S. und Sturdivant, R. (2013). **Applied logistic regression**. Wiley.

Illig, M. (2016). **Das Ausmaß der cerebralen Mikroangiopathie als Einflussgröße für akutes und endgültiges Schlaganfallvolumen**. Medizinische Fakultät der Universität Hamburg.

Institute for Health Metrics and Evaluation (University of Washington). (2021). **GBD Results**. http://ghdx.healthdata.org/gbd-results-tool.

Jayaraman, M., McTaggart, R. und Goyal, M. (2017). **Unresolved issues in thrombectomy**. *Current Neurology and Neuroscience Reports*, 17(9):69.

Jenkinson, M., Beckmann, C., Behrens, T., Woolrich, M. und Smith, S. (2012). **FSL**. *Neuroimage*, 62(2):782–90.

Johnson, C., Nguyen, M., Roth, G., Nichols, E., Alam, T., Abate, D., Abd-Allah, F., Abdelalim, A., Abraha, H., Abu-Rmeileh, N., Adebayo, O., Adeoye, A., Agarwal, G., Agrawal, S., Aichour, A., Aichour, I., Aichour, M., … Murray, C. (2019). **Global, regional, and national burden of stroke, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016**. *The Lancet Neurology*, 18(5):439–58.

Jonsdottir, K.Y., Østergaard, L. und Mouridsen, K. (2009). **Predicting tissue outcome from acute stroke magnetic resonance imaging: improving model performance by optimal sampling of training data**. *Stroke*, 40(9):3006–11.

Kammen, M. Van, Moomaw, C., Schaaf, I. Van Der, Brown, R., Woo, D., Broderick, J., Mackey, J., Rinkel, G.J., Huston, J. und Ruigrok, Y. (2018). **Heritability of circle of Willis variations in families with intracranial aneurysms**. *PLoS ONE*, 13(1):e0191974.

Katan, M. und Luft, A. (2018). **Global Burden of Stroke**. *Seminars in Neurology*, 38(2):208–11.

Kellner, E. (2014). **Quantitative Bestimmung der cerebralen Durchblutung mittels dynamischer Suszeptibilitätskontrast-Magnetresonanztomographie**. Albert-Ludwigs-Universität Freiburg.

Kelly-Hayes, M. (2010). **Influence of age and health behaviors on stroke risk: Lessons from longitudinal studies**. *Journal of the American Geriatrics Society*, 58(s2):325–28.

Kemmling, A., Flottmann, F., Forkert, N.D., Minnerup, J., Heindel, W., Thomalla, G., Eckert, B., Knauth, M., Psychogios, M., Langner, S. und Fiehler, J. (2015). **Multivariate dynamic prediction of ischemic infarction and tissue salvage as a function of time and degree of recanalization**. *Journal of Cerebral Blood Flow & Metabolism*, 35(9):1397–1405.

Kim, B., Kang, H., Kim, H.-J., Ahn, S.-H., Kim, N., Warach, S. und Kang, D.-W. (2014). **Magnetic resonance imaging in acute ischemic stroke treatment**. *Journal of Stroke*, 16(3):131.

Kim, J.-T., Cho, B.-H., Choi, K.-H., Park, M.-S., Kim, B., Park, J.-M., Kang, K., Lee, S., Kim, J., Cha, J.-K., Kim, D.-H., Nah, H.-W., Park, T., Park, S.-S., Lee, K., Lee, J., Hong, K.-S., … Cho, K.-H. (2019). **Magnetic resonance imaging versus computed tomography angiography based selection for endovascular therapy in patients with acute ischemic stroke**. *Stroke*, 50(2):365–72.

Kleindorfer, D., Miller, R., Moomaw, C., Alwell, K., Broderick, J., Khoury, J., Woo, D., Flaherty, M., Zakaria, T. und Kissela, B. (2007). **Designing a message for public education regarding stroke: Does FAST capture enough stroke?** *Stroke*, 38(10):2864–68.

Klug, J., Dirren, E., Preti, M., Machi, P., Kleinschmidt, A., Vargas, M., Ville, D. Van De und Carrera, E. (2020). **Integrating regional perfusion CT information to improve prediction of infarction after stroke**. *Journal of Cerebral Blood Flow & Metabolism*, 41(3):502–10.

Krishnamurthi, R., Ikeda, T. und Feigin, V. (2020). **Global, regional and country-specific burden of ischaemic stroke, intracerebral haemorrhage and subarachnoid haemorrhage: A systematic analysis of the Global Burden of Disease Study 2017**. *Neuroepidemiology*, 54(2):171–79.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M. und Malohlava, M. (2018). **h2o: R Interface for H2O**. https://github.com/h2oai/h2o-3.

Li, W., Yin, Y., Quan, X. und Zhang, H. (2019). **Gene expression value prediction based on XGBoost algorithm**. *Frontiers in Genetics*, 10:1077.

Liaw, A. und Wiener, M. (2002). **Classification and regression by randomForest**. *R News*, 2(3):18–22. https://cran.r-project.org/doc/rnews/.

Livne, M., Boldsen, J., Mikkelsen, I., Fiebach, J., Sobesky, J. und Mouridsen, K. (2018). **Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke**. *Stroke*, 49(4):912–18.

Luengo-Fernandez, R., Violato, M., Candio, P. und Leal, J. (2019). **Economic burden of stroke across Europe : A population-based cost analysis**.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. und Lee, S.-I. (2020). **From local explanations to global understanding with explainable AI for trees**. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S., Erion, G. und Lee, S.-I. (2019). **Consistent individualized feature attribution for tree ensembles**. https://arxiv.org/abs/1802.03888v3.

Lundberg, S. und Lee, S.-I. (2017). **A unified approach to interpreting model predictions**. http://arxiv.org/abs/1705.07874.

Lyden, P. (2017). **Using the National Institutes of Health Stroke Scale: A cautionary tale**. *Stroke*, 48(2):513–19.

Maier, O., Schröder, C., Forkert, N., Martinetz, T. und Handels, H. (2015). **Classifiers for ischemic stroke lesion segmentation: A comparison study**. *PLoS ONE*, 10(12):e0145118.

Mainali, S., Darsie, M. und Smetana, K. (2021). **Machine learning in action: Stroke diagnosis and outcome prediction**. *Frontiers in Neurology*, 12.

McKinley, R., Häni, L., Gralla, J., El-Koussy, M., Bauer, S., Arnold, M., Fischer, U., Jung, S., Mattmann, K., Reyes, M. und Wiest, R. (2016). **Fully automated stroke tissue estimation using random forest classifiers (FASTER)**. *Journal of Cerebral Blood Flow & Metabolism*, 37(8):2728–41.

Mckinley, R., Levin, H., Wiest, R. und Reyes, M. (2015). **Segmenting the ischemic penumbra: A decision forest approach with automatic threshold finding**. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing*, 275–83.

Michel, P., Odier, C., Rutgers, M., Reichhart, M., Maeder, P., Meuli, R., Wintermark, M., Maghraoui, A., Faouzi, M., Croquelois, A. und Ntaios, G. (2010). **The Acute STroke Registry and Analysis of Lausanne (ASTRAL)**. *Stroke*, 41(11):2491–98.

Mirexon. o. J. **Voxel human brain**. https://www.dreamstime.com/stock-illustration-voxel-human-brain-d-render-isolated-white-background-image51834849.

Molnar, C. (2019). **Interpretable Machine Learning**.

Moseley, M., Cohen, Y., Mintorovitch, J., Chileuitt, L., Shimizu, H., Kucharczyk, J., Wendland, M. und Weinstein, P. (1990). **Early detection of regional cerebral ischemia in cats: Comparison of diffusion- and T2-weighted MRI and spectroscopy**. *Magnetic Resonance in Medicine*, 14(2):330–46.

Mouridsen, K., Thurner, P. und Zaharchuk, G. (2020). **Artificial intelligence applications in stroke**. *Stroke*, 51(8):2573–79.

Neil, J. (1997). **Measurement of water motion (apparent diffusion) in biological systems**. *Concepts in Magnetic Resonance*, 9(6):385–401.

Nguyen, V., Pien, H. und Menenzes, N. (2008). **Stroke tissue outcome prediction using a spatially-correlated model**. *Pan Pacific Imaging Conference*, 3:238–41.

Nielsen, A., Hansen, M., Tietze, A. und Mouridsen, K. (2018). **Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning**. *Stroke*, 49(6):1394–1401.

Nielsen, D. (2016). **Tree Boosting With XGBoost**. Norwegian University of Science and Technologie. https://ntnuopen.ntnu.no/ntnuxmlui/bitstream/handle/11250/2433761/16128_fulltext.pdf.

Ortiz, G. und Sacco, R. (2007). **National institutes of health stroke scale (NIHSS)**. *Wiley Encyclopedia of Clinical Trials*, 1–9.

Ozenne, B., Cho, T., Mikkelsen, I., Hermier, M., Ribe, L., Thomalla, G., Pedraza, S., Baron, J., Roy, P., Berthezène, Y., Nighoghossian, N., Østergaard, L. und Maucort-Boulch, D. (2015). **Evaluation of early reperfusion criteria in acute ischemic stroke**. *Journal of Neuroimaging*, 25(6):952–58.

Parliament, E. (2016). **Regulation (EU) 2016/679 of the the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR)**. *Official Journal of the European Union*, 59(119):1–88.

Payabvash, S., Souza, L., Wang, Y., Schaefer, P.W., Furie, K., Halpern, E., Gonzalez, R. und Lev, M. (2011). **Regional ischemic vulnerability of the brain to hypoperfusion: The need for location specific computed tomography perfusion thresholds in acute stroke patients**. *Stroke*, 42(5):1255–60.

Pinto, A., McKinley, R., Alves, V., Wiest, R., Silva, C. und Reyes, M. (2018). **Stroke lesion outcome prediction based on MRI imaging combined with clinical information**. *Frontiers in Neurology*, 9:1060.

Pinto, A., Pereira, S., Meier, R., Alves, V., Wiest, R., Silva, C. und Reyes, M. (2018). **Enhancing clinical MRI perfusion maps with data-driven maps of complementary nature for lesion outcome prediction**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11072 LNCS:107–15.

Probst, P., Wright, M. und Boulesteix, A.-L. (2019). **Hyperparameters and tuning strategies for random forest**. *WIREs Data Mining and Knowledge Discovery*, 9(3):1–15.

Purushotham, A., Campbell, B., Straka, M., Mlynash, M., Olivot, J.-M., Bammer, R., Kemp, S., Albers, G. und Lansberg, M. (2013). **Apparent diffusion coefficient threshold for delineation of ischemic core**. *International Journal of Stroke*, 10(3):348–53.

Quinn, T., Dawson, J., Walters, M. und Lees, K. (2009). **Reliability of the modified rankin scale: A systematic review**. *Stroke*, 40(10):3393–95.

Reich, A. und Nikoubashman, O. (2016). **Ischämischer Schlaganfall (zerebrale Ischämie)**. In *Neurologische Notfälle*, 1–23. Springer Berlin Heidelberg.

Rekik, I., Allassonnière, S., Carpenter, T. und Wardlaw, J. (2012). **Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal**. *NeuroImage: Clinical*, 1(1):164–78.

Revolution Analytics und Weston, S. (2015). **foreach: Provides foreach looping construct for R**. https://cran.r-project.org/package=foreach.

Ribeiro, M., 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*, 1135–44. ACM.

Riedel, C., Zimmermann, P., Jensen-Kondering, U., Stingele, R., Deuschl, G. und Jansen, O. (2011). **The importance of size: Successful recanalization by intravenous thrombolysis in acute anterior stroke depends on thrombus length**. *Stroke*, 42(6):1775–77.

Ringleb, P., Köhrmann, M. und Jansen, O. (2021). **Akuttherapie des ischämischen Schlaganfalls, S2e-Leitlinie**. Deutsche Gesellschaft für Neurologie.

Robben, D., Christiaens, D., Rangarajan, J., Gelderblom, J., Joris, P., Maes, F. und Suetens, P. (2016). **A voxel-wise, cascaded classification approach to ischemic stroke lesion segmentation**. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 254–65. Springer International Publishing.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. und Müller, M. (2011). **pROC: an open-source package for R and S+ to analyze and compare ROC curves**. *BMC Bioinformatics*, 12:77.

Ronneberger, O., Fischer, P. und Brox, T. (2015). **U-Net: Convolutional networks for biomedical image segmentation**. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–41. Springer International Publishing.

Saarinen, J., Rusanen, H. und Sillanpää, N. (2014). **Collateral score complements clot location in predicting the outcome of intravenous thrombolysis**. *American Journal of Neuroradiology*, 35(10):1892–96.

Saarinen, J., Sillanpää, N., Rusanen, H., Hakomäki, J., Huhtala, H., Lähteelä, A., Dastidar, P., Soimakallio, S. und Elovaara, I. (2012). **The mid-M1 segment of the middle cerebral artery is a cutoff clot location for good outcome in intravenous thrombolysis**. *European Journal of Neurology*, 19(8):1121–27.

Sacco, R., Kasner, S., Broderick, J., Caplan, L., Connors, J., Culebras, A., Elkind, M.S., George, M., Hamdan, A., Higashida, R., Hoh, B., Janis, L., Kase, C., Kleindorfer, D., Lee, J., Moseley, M., Peterson, E., … Vinters, H. (2013). **An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American Heart Association/American Stroke Association**. *Stroke*, 44(7):2064–89.

Scalzo, F., Hao, Q., Alger, J., Hu, X. und Liebeskind, D. (2012). **Regional prediction of tissue fate in acute ischemic stroke**. *Annals of Biomedical Engineering*, 40(10):2177–87.

Schellinger, P., Bryan, R., Caplan, L., Detre, J., Edelman, R., Jaigobin, C., Kidwell, C., Mohr, J., Sloan, M., Sorensen, A. und Warach, S. (2010). **Evidence-based guideline: The role of diffusion and perfusion MRI for the diagnosis of acute ischemic stroke (RETIRED)**. *Neurology*, 75(2):177–85.

Schlaug, G., Benfield, A., Baird, A., Siewert, B., Lövblad, K., Parker, R., Edelman, R. und Warach, S. (1999). **The ischemic penumbra: Operationally defined by diffusion and perfusion MRI**. *Neurology*, 53(7):1528–37.

Schubert, F. und Lalouschek, W. (2011). **Schlaganfall**. In *Klinische Neuropsychologie*, 345–56. Springer Vienna.

Sesto, F. (2013). **Koronare Herzkrankheit II – Fragen – Antworten**. Berlin Heidelberg New York: Springer-Verlag.

Shapley, L. (1953). **A value for n-person games**. *Contributions to the Theory of Games (AM-28)*, 2:307–18.

Shen, Q. und Duong, T. (2008). **Quantitative prediction of ischemic stroke tissue fate**. *NMR in Biomedicine*, 21(8):839–48.

Sheth, S. und Liebeskind, D. (2013). **Imaging evaluation of collaterals in the brain: Physiology and clinical translation**. *Current Radiology Reports*, 2(1):29.

Siemens Medical. (2003). **Magnete, Spins und Resonanzen – Eine Einführung in die Grundlagen der Magnetresonanztomographie**. *Siemens Medical*, 1–238.

Siemonsen, S., Forkert, N., Hansen, A., Kemmling, A., Thomalla, G. und Fiehler, J. (2014). **Spatial distribution of perfusion abnormality in acute MCA occlusion is associated with likelihood of later recanalization**. *Journal of Cerebral Blood Flow and Metabolism*, 34(5):813–19.

Sohn, C.-H., Sohn, S.-I., Chang, H.-W. und Demchuk, A. (2007). **Postcontrast time-of-flight MR angiography demonstrating collateral flow in acute stroke**. *Stroke*, 38(4):1132.

Sotoudeh, H., Bag, A. und Brooks, M. (2019). **“Code-Stroke” CT perfusion; challenges and pitfalls**. *Academic Radiology*, 26(11):1565–79.

Sourbron, S. (2010). **Technical aspects of MR perfusion**. *European Journal of Radiology*, 76(3):304–13.

Swieten, J. Van, Koudstaal, P., Visser, M., Schouten, H. und Gijn, J. Van. (1988). **Interobserver agreement for the assessment of handicap in stroke patients**. *Stroke*, 19(5):604–7.

Tawil, S. El und Muir, K. (2017). **Thrombolysis and thrombectomy for acute ischaemic stroke**. *Clinical Medicine, Journal of the Royal College of Physicians of London*, 17(2):161–65.

Thomalla, G., Audebert, H., Berger, K., Fiebach, J., Fiehler, J., Kaps, M., Neumann-Haefelin, T., Schellinger, P., Siebler, M., Sobesky, J., Villringer, A., Witte, O. und Röther, J. (2009). **Bildgebung beim Schlaganfall – eine Übersicht und Empfehlungen des Kompetenznetzes Schlaganfall**. *Aktuelle Neurologie*, 36(07):354–67.

Tong, D., Reeves, M., Hernandez, A., Zhao, X., Olson, D., Fonarow, G., Schwamm, L. und Smith, E. (2012). **Times from symptom onset to hospital arrival in the get with the guidelines-stroke program 2002 to 2009**. *Stroke*, 43(7):1912–17.

Truelsen, T., Piechowski-Jóźwiak, B., Bonita, R., Mathers, C., Bogousslavsky, J. und Boysen, G. (2006). **Stroke incidence and prevalence in Europe: A review of available data**. *European Journal of Neurology*, 13(6):581–98.

Vavilala, M., Lee, L. und Lam, A. (2002). **Cerebral blood flow and vascular physiology**. *Anesthesiology Clinics of North America*, 20(2):247–64.

Vert, C., Parra-Fariñas, C. und Rovira, À. (2017). **MR imaging in hyperacute ischemic stroke**. *European Journal of Radiology*, 96:125–32.

Vilela, P. und Rowley, H. (2017). **Brain ischemia: CT and MRI techniques in acute ischemic stroke**. *European Journal of Radiology*, 96(11):162–72.

Wang, F., Kaushal, R. und Khullar, D. (2019). **Should health care demand interpretable artificial intelligence or accept “black box” medicine?** *Annals of Internal Medicine*, 172(1):59.

Warfield, S., Zou, K. und Wells, W. (2004). **Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation**. *IEEE Transactions on Medical Imaging*, 23(7):903–21.

Wickham, H. (2016). **ggplot2 - Elegant graphics for data analysis**. Cham, Heidelberg, New York, Dordrecht, London: Springer International Publishing.

Wickham, H., François, R., Henry, L. und Müller, K. (2020). **dplyr: A grammar of data manipulation**. https://github.com/tidyverse/dplyr.

Winder, A., Siemonsen, S., Flottmann, F., Thomalla, G., Fiehler, J. und Forkert, N. (2019). **Technical considerations of multi-parametric tissue outcome prediction methods in acute ischemic stroke patients**. *Scientific Reports*, 9(1):13208.

Winder, A., Wilms, M., Fiehler, J. und Forkert, N. (2021). **Treatment efficacy analysis in acute ischemic stroke patients using in silico modeling based on machine learning: A proof-of-principle**. *Biomedicines*, 9(10):1357.

Winzeck, S., Hakim, A., McKinley, R., Pinto, J., Alves, V., Silva, C., Pisov, M., Krivov, E., Belyaev, M., Monteiro, M., Oliveira, A., Choi, Y., Paik, M., Kwon, Y., Lee, H., Kim, B., Won, J., … Reyes, M. (2018). **ISLES 2016 and 2017 – benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI**. *Frontiers in Neurology*, 9:679.

Woo, D., Broderick, J., Kothari, R., Lu, M., Brott, T., Lyden, P., Marler, J. und Grotta, J. (1999). **Does the National Institutes of Health Stroke Scale favor left hemisphere strokes?** *Stroke*, 30(11):2355–59.

Wouters, A., Christensen, S., Straka, M., Mlynash, M., Liggins, J., Bammer, R., Thijs, V., Lemmens, R., Albers, G. und Lansberg, M. (2017). **A comparison of relative time to peak and tmax for mismatch-based patient selection**. *Frontiers in Neurology*, 8:539.

Wu, O., Christensen, S., Hjort, N., Dijkhuizen, R., Kucinski, T., Fiehler, J., Thomalla, G., Röther, J. und Østergaard, L. (2006). **Characterizing physiological heterogeneity of infarction risk in acute human ischaemic stroke using MRI**. *Brain*, 129(9):2384–93.

Wu, O., Koroshetz, W., Østergaard, L., Buonanno, F., Copen, W., Gonzalez, R., Rordorf, G., Rosen, B., Schwamm, L., Weisskoff, R. und Sorensen, A. (2001). **Predicting tissue outcome in acute human cerebral ischemia using combined diffusion- and perfusion-weighted MR imaging**. *Stroke*, 32(4):933–42.

Wu, O., Østergaard, L., Weisskoff, R., Benner, T., Rosen, B. und Sorensen, A. (2003). **Tracer arrival timing-insensitive technique for estimating flow in MR perfusion-weighted imaging using singular value decomposition with a block-circulant deconvolution matrix**. *Magnetic Resonance in Medicine*, 50(1):164–74.

Yilmaz, U. (2015). **Diffusionsgewichtete Bildgebung bei akutem Schlaganfall**. *Der Radiologe*, 55(9):771–74.

Yu, Y., Xie, Y., Thamm, T., Gong, E., Ouyang, J., Huang, C., Christensen, S., Marks, M., Lansberg, M., Albers, G. und Zaharchuk, G. (2020). **Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging**. *JAMA Network Open*, 3(3):1–13.

Zaro-Weber, O., Moeller-Hartmann, W., Siegmund, D., Kandziora, A., Schuster, A., Heiss, W.-D. und Sobesky, J. (2016). **MRI-based mismatch detection in acute ischemic stroke: Optimal PWI maps and thresholds validated with PET**. *Journal of Cerebral Blood Flow & Metabolism*, 37(9):3176–83.

Zihni, E., Madai, V., Livne, M., Galinovic, I., Khalil, A., Fiebach, J. und Frey, D. (2020). **Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome**. *PLoS ONE*, 15(4):e0231166.

Zihni, E., McGarry, B. und Kelleher, J. (2021). **An analysis of the interpretability of neural networks trained on magnetic resonance imaging for stroke outcome prediction**. In *Proceedings of the 29th International Society for Magnetic Resonance in Medicine (ISMRM) Annual Scientific Meeting and Exhibition*, 3503.

Zoppo, G. Del, Saver, J., Jauch, E. und Adams, H. (2009). **Expansion of the time window for treatment of acute ischemic stroke with intravenous tissue plasminogen activator: A science advisory from the American heart association/american stroke association**. *Stroke*, 40(8):2945–48.

Hinweise zur Veröffentlichung

Teile dieser Arbeit wurden veröffentlicht in:

1) Flottmann, F., Broocks, G., Faizy, T., Ernst, M., Forkert, N., Grosser, M., Thomalla, G., Siemonsen, S., Fiehler, J. und Kemmling, A. (2017). **CT-perfusion stroke imaging: A threshold free probabilistic approach to predict infarct volume compared to traditional ischemic thresholds**. *Scientific Reports*, 7(1):6679.

2) Grosser, M., Gellißen, S., Borchert, P., Sedlacik, J., Nawabi, J., Fiehler, J. und Forkert, N. (2020a). **Improved multi-parametric prediction of tissue outcome in acute ischemic stroke patients using spatial features**. *PLoS ONE*, 15(1):e0230653.

3) Grosser, M., Gellißen, S., Borchert, P., Sedlacik, J., Nawabi, J., Fiehler, J. und Forkert, N. (2020b). **Localized prediction of tissue outcome in acute ischemic stroke patients using diffusion- and perfusion-weighted MRI datasets**. *PLoS ONE*, 15(11):e0241917.

Danksagung

Ich möchte all jenen, die auf unterschiedliche Weise zu dieser Arbeit beigetragen haben und ohne die diese Arbeit nicht möglich gewesen wäre, meinen Dank aussprechen.

Zuallererst bedanke ich mich herzlich bei meinem Doktorvater Prof. Dr. Jens Fiehler, der mir die Möglichkeit gab, mich an seinem Institut mit diesem spannenden Thema näher auseinanderzusetzen und mich fortwährend in vielen Gesprächen und in allen Belangen dieser Arbeit unterstützte.

Ein herzliches Dankeschön gilt auch Prof. Dr. Karl Wegscheider und Prof. Dr. Jens Struckmeier, die mir als Zweit- und Drittgutachter bei unseren regelmäßigen Betreuertreffen immer wieder gute Ratschläge und kritische Denkanstöße mit auf den Weg gaben.

Ein ganz besonderer Dank gilt Prof. Dr. Nils Forkert, der einen Großteil der fachlichen Betreuung übernahm und sich trotz verschiedener Zeitzonen in unseren unzähligen Skype-Gesprächen stets geduldig die Zeit nahm, meine fachlichen Fragen zu beantworten. Darüber hinaus leistete er entscheidende Beiträge zum Schreiben und Einreichen unserer beiden Veröffentlichungen.

Ein besonders großer Dank gebührt ebenfalls PD. Dr. Susanne Gellißen, die mich vor allem während der ersten Hälfte dieser Dissertation intensiv betreute und deren Tür immer für jegliche Fragen offenstand. Insbesondere beantwortete sie mir alle Fragen zu den I-KNOW-Daten und ließ mich auch sonst stets auf verständliche Weise an ihrer Expertise in der Schlaganfalldiagnostik teilhaben.

Ein großer Dank gilt Patrick Borchert, der mir bei der Einarbeitung in die Bildverarbeitungssoftware tatkräftig und unermüdlich zur Seite stand und dabei mit seiner direkten und humorvollen Art schnell zu einem engen Freund wurde.

Ein großes Dankeschön geht auch an PD. Dr. Jan Sedlacik, der immer ein offenes Ohr für mich hatte und für allerlei technische Probleme stets eine pragmatische Lösung fand.

Ich möchte mich bei allen Kollegen am Institut für die gute und ergiebige Zusammenarbeit bedanken. Insbesondere danke ich PD. Dr. Fabian Flottmann, PD. Dr. Tobias Faizy, Dr. Jawed Nawabi, Dr. Tanja Schneider und Dr. Jan-Niklas Hochstein.

Darüber hinaus bedanke ich mich bei Fabian Temme, durch dessen Forschungsprojekt die Stelle für diese Promotion ermöglicht wurde.

Ich danke all meinen Freunden, die mich während dieser Promotion begleitet haben. Vor allem danke ich Alexander Achner, Robert Banas, Mahkameh Amini, Thomas Rosario, Muhammad Jimmy Kamboh, Henning Bumann, Amanuel Woldeyohannes, Michael Stanko und Dr. Sebastian Bannasch für ihren Zuspruch und ihre aufmunternde Unterstützung.

Ich möchte mich von ganzem Herzen bei meiner Familie bedanken. Insbesondere bei meiner Mutter, Christa Grosser, die mich in allen Lebenslagen unterstützt und mir und meiner Schwester immer zur Seite steht.

Ein liebevolles Dankeschön gilt meiner Frau Elena Grosser für ihre Geduld und ihre fürsorgliche Unterstützung.

Curriculum Vitae

Lebenslauf aus datenschutzrechtlichen Gründen nicht enthalten

Eidesstattliche Erklärung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.

Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.

Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Unterschrift: .....