

Universitätsklinikum Hamburg Eppendorf

Klinik und Poliklinik für Neuroradiologische Diagnostik und Intervention

Klinikdirektor

Prof. Dr. med. Jens Fiehler

Automatisierte Quantifizierung der Revaskularisierung nach mechanischer Thrombektomie beim ischämischen Schlaganfall mittels neuronaler Netzwerke

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
an der Medizinischen Fakultät der Universität Hamburg.

Vorgelegt von

Moritz Waldmann
aus Hildesheim

Hamburg 2021

Angenommen von der Medizinischen Fakultät am: **05.07.2022**

Veröffentlicht mit Genehmigung der Medizinischen Fakultät der Universität Hamburg

Prüfungsausschuss, der/die Vorsitzende: **Prof. Dr. Götz Thomalla**

Prüfungsausschuss, 2. Gutachter/in: **PD Dr. Fabian Flottmann**

Abkürzungsverzeichnis

ACA	A. cerebri anterior
ACC	<i>Accuracy</i>
ACC_{±1}	<i>Sliding window accuracy</i>
ACE	A. carotis externa
ACI	A. carotis interna
ACM	A. cerebri media
Acom	A. communicans anterior
AIS	Akuter ischämischer Schlaganfall
ATA	A. temporalis anterior
C_κ	Cohens κ
CNN	Konvolutionsnetz (<i>convolutional neural network</i>)
DL	<i>Deep learning</i>
DSA	Digitale Subtraktionsangiographie
FDA	<i>Food and drug administration</i>
FOV	<i>Field of view</i>
GRU	<i>Gated recurrent unit</i>
i.v.	Intravenös
MT	Mechanische Thrombektomie
NN	Neuronales Netzwerk
Pcom	A. communicans posterior
ps	Perfusionsscore
TICI	<i>Thrombolysis in cerebral infarction</i>

Inhaltsverzeichnis

Abkürzungsverzeichnis	I
Inhaltsverzeichnis	II
1 Originalpublikation	1
2 Einleitung	17
2.1 Behandlung des ischämischen Schlaganfalls	17
2.2 Der TICI Score	18
2.3 Untersuchung zur Interrater-Reliabilität bei der Bestimmung des TICI Scores	21
2.4 <i>Deep learning</i> Ansätze in der medizinischen Bildverarbeitung	24
2.5 Training neuronaler Netzwerke	25
3 Zielsetzung	27
4 Material und Methoden	29
4.1 Erhebung des Datensatzes	29
4.2 Ein- und Ausschlusskriterien zur Aufnahme in den Datensatz	29
4.3 Externer Datensatz	32
4.4 Training und Testung der neuronalen Netzwerke	32
4.5 Bestimmung der Perfusionsscores	33
4.7 Statistische Auswertung der Interrater-Reliabilität für die Perfusionsscores	36
5 Ergebnisse	37
5.1 Hypothese 1: Eignung neuronaler Netzwerke zur automatischen Quantifizierung des Behandlungserfolgs der MT nach TICI	37
5.1.1 Erhebung des <i>in-house</i> Datensatzes	37
5.1.2 Externer Datensatz: Bestimmung von Intra- und Interrater-Reliabilität	37
5.1.3 Statistische Analyse der Prädiktionsgenauigkeit der NN auf die verschiedenen Datensätze	38
5.1.4 Einordnung der Prädiktionsgenauigkeit mit in der Literatur beschriebenen Werten zur Interrater-Reliabilität	41

5.2	Hypothese 2: Perfusionsscores als Alternative zum TICI Score	42
5.2.1	Bestimmung der Perfusionsscores	42
5.2.2	Bestimmung der Interrater-Reliabilität.....	46
6	Diskussion.....	47
6.1	Hypothese 1: Eignung neuronaler Netzwerke zur automatischen Quantifizierung des Behandlungserfolgs der MT nach TICI.....	47
6.2	Hypothese 2: Perfusionsscores als zum TICI Score potentiell alternatives Label	49
6.2.1	Bestimmung der Perfusionsscores	49
6.2.2	Interrater-Reliabilität der Perfusionsscores	50
7	Zusammenfassung.....	52
8	Summary.....	53
9	Literatur	54
10	Erklärung des Eigenanteils an der Publikation	57
11	Danksagung	58
12	Lebenslauf	59
13	Eidesstattliche Versicherung	60

1 Originalpublikation

Wesentliche Teile der vorliegenden Arbeit sind in der vorliegenden Publikation beschrieben.
Eine detailliertere Einordnung der Publikation findet sich in den anschließenden Kapiteln.

CLINICAL AND POPULATION SCIENCES

Deep Learning–Based Automated Thrombolysis in Cerebral Infarction Scoring

A Timely Proof-of-Principle Study

Maximilian Nielsen¹, MSc*; Moritz Waldmann, PhD*; Andreas M. Frölich², MD; Fabian Flottmann³, MD; Evelin Hristova, MSc; Martin Bendszus, MD; Fatih Seker⁴, MD; Jens Fiehler⁵, MD†; Thilo Sentker⁶, PhD†; Rene Werner¹, PhD†

BACKGROUND AND PURPOSE: Mechanical thrombectomy is an established procedure for treatment of acute ischemic stroke. Mechanical thrombectomy success is commonly assessed by the Thrombolysis in Cerebral Infarction (TICI) score, assigned by visual inspection of X-ray digital subtraction angiography data. However, expert-based TICI scoring is highly observer-dependent. This represents a major obstacle for mechanical thrombectomy outcome comparison in, for instance, multicentric clinical studies. Focusing on occlusions of the M1 segment of the middle cerebral artery, the present study aimed to develop a deep learning (DL) solution to automated and, therefore, objective TICI scoring, to evaluate the agreement of DL- and expert-based scoring, and to compare corresponding numbers to published scoring variability of clinical experts.

METHODS: The study comprises 2 independent datasets. For DL system training and initial evaluation, an in-house dataset of 491 digital subtraction angiography series and modified TICI scores of 236 patients with M1 occlusions was collected. To test the model generalization capability, an independent external dataset with 95 digital subtraction angiography series was analyzed. Characteristics of the DL system were modeling TICI scoring as ordinal regression, explicit consideration of the temporal image information, integration of physiological knowledge, and modeling of inherent TICI scoring uncertainties.

RESULTS: For the in-house dataset, the DL system yields Cohen's kappa, overall accuracy, and specific agreement values of 0.61, 71%, and 63% to 84%, respectively, compared with the gold standard: the expert rating. Values slightly drop to 0.52/64%/43% to 87% when the model is, without changes, applied to the external dataset. After model updating, they increase to 0.65/74%/60% to 90%. Literature Cohen's kappa values for expert-based TICI scoring agreement are in the order of 0.6.

CONCLUSIONS: The agreement of DL- and expert-based modified TICI scores in the range of published interobserver variability of clinical experts highlights the potential of the proposed DL solution to automated TICI scoring.

GRAPHIC ABSTRACT: An online [graphic abstract](#) is available for this article.

Key Words: angiography, digital subtraction ■ deep learning ■ ischemic stroke ■ middle cerebral artery ■ thrombectomy

In the past few years, several breakthrough trials have shown an increased therapeutic efficacy of mechanical thrombectomy (MT) over intravenous fibrinolysis for treatment of patients with acute ischemic stroke (AIS) due to the occlusion of large intracranial vessels.^{1–4} Endovascular revascularization is now considered standard therapy in patients with AIS with large vessel occlusions.^{5–8} With the growing experience in performing

MT, the indication for MT treatment and the number of treated patients is continuously expanding.^{9,10}

See related article, p 3505

MT success is usually measured using an ordinal scale, the Thrombolysis in Cerebral Infarction (TICI) score.¹¹

Correspondence to: René Werner, PhD, University Medical Center, Hamburg-Eppendorf, Martinistr 52, 20246 Hamburg, Germany. Email r.werner@uke.de

*M. Nielsen and M. Waldmann contributed equally.

†J. Fiehler, T. Sentker, and R. Werner contributed equally.

The Data Supplement is available with this article at <https://www.ahajournals.org/doi/suppl/10.1161/STROKEAHA.120.033807>.

For Sources of Funding and Disclosures, see page 3503.

© 2021 American Heart Association, Inc.

Stroke is available at www.ahajournals.org/journal/str

Nonstandard Abbreviations and Acronyms

AIS	acute ischemic stroke
DL	deep learning
DSA	digital subtraction angiography
MT	mechanical thrombectomy
SA	specific agreement
TICI	Thrombolysis in Cerebral Infarction
mTICI	modified TICI

The rating is determined based on the perfusion visible in time-resolved X-ray digital subtraction angiography (DSA) image series acquired during the intervention.^{12,13} Despite its common use and the coarse distribution of the categories of the original TICI scale and suggested modifications (modified TICI [mTICI]),¹⁴ expanded TICI¹³, the inter- and intraobserver variability of TICI scoring is surprisingly high, with corresponding numbers varying from study to study (Table 1).^{13,15–19}

Respective scores are, nevertheless, used for evaluation purposes in clinical studies and as performance metric in hospitals performing MT. Given the high interobserver variability, it remains questionable whether such scores and study results are comparable and suited for the envisioned purpose if they are assigned by different physicians and centers. Even if TICI scoring is performed by specialized CoreLabs, scoring variability of the CoreLab experts remains challenging. Moreover, a recent study by Zhang et al²⁰ highlights pronounced discrepancies between CoreLab-assigned TICI scores and those assigned by local clinicians directly after the procedure. Thus, there is the need for an objective, that is, automatic and deterministic TICI scoring for MT, as it would address the problem of limited comparability between studies.²¹ However, even in the current age of deep learning (DL) where neural networks offer a way to reliably classify complex data, algorithms capable of exploiting the inherent information of temporal DSA series have been rarely reported on as of yet. In the context of TICI scoring, the few existing approaches, although promising, require a larger amount of user interaction (expert-based cropping and normalization of the DSA series,²² delineation of relevant areas²³), rendering them impractical for clinical application and still dependent on operator-specific input. At the moment, machine learning in AIS rather focuses on aspects like outcome prediction,^{24,25} detection of ischemia,²⁶ and hemorrhage²⁷ or thrombus localization²⁸ using computed tomography or magnetic resonance imaging data.²⁹ Reasons for the sparse usage of DSA image data in machine learning might be due to the complex nature of angiographic runs. Further, the low standardization between institutions in acquiring DSA series leads to a high variability in, for example, length

of series, applied amount of contrast agents, projection angles, and magnification factors in available datasets.

Addressing these challenges, the present work describes a first DL-based approach to fully automatically assign TICI scores to given DSA series. Key aspects of the DL system development were modeling of TICI scoring as an ordinal regression problem and taking into account human rater scoring uncertainty; explicit consideration of the spatio-temporal DSA image information during system training and application; and integration of straight-forward physiological prior knowledge about the perfusion process. Our proof-of-concept study focuses on M1 occlusions, the most frequent large vessel occlusion site.³⁰ Thereby, we circumvent ambiguities regarding the definition of the initial occlusion considered for TICI scoring. The model performance and its capability to generalize to unseen data are investigated by means of independent DSA datasets from 2 centers and using the mTICI scale.

METHODS

Data that support the findings of this study are available from the corresponding author upon reasonable request. To foster reuse, the source code and the final DL network are available at <https://github.com/IPMI-ICNS-UKE/DeepTICI>. A MI-CLAIM checklist³¹ is provided as [Data Supplement](#).

Patients

The study is based on 2 independent datasets: an in-house dataset and an external dataset.

For the in-house dataset, DSA series of patients with AIS treated by MT between 2010 and 2018 at the University Medical Center Hamburg-Eppendorf were retrospectively selected from the clinical database and anonymized. MT was performed by experienced neuroradiologists (>5 years experience). Inclusion criteria for the study were (1) only M1 occlusion, (2) availability of bi-plane projection data (lateral and frontal, following mTICI scoring recommendations¹⁴), (3) a minimum DSA series length of 5 frames, (4) >65% of the middle cerebral artery territory visible and >30% of the field of view representing the middle cerebral artery territory, (5) no radiopaque foreign body projections over the middle cerebral artery territory, and (6) no further pathology (eg, dissection, subarachnoid hemorrhage, aneurysm). To reflect clinical variability, DSA series with retrograde perfusion, contralateral contrast filling, moderate motion artefacts or perfusion of the posterior territory via the posterior communicating artery were not excluded. Exemplary DSA movies are provided as [Data Supplement](#).

From a total of 789 evaluated patient datasets, 491 DSA series of 236 patients matched the inclusion criteria (main exclusion reason: different occlusion site than M1, 418/789 patients; see details and flow chart in the [Data Supplement](#)). When >2 DSA series per patient were included, the additional series reflected the situation after intermediate MT runs. The assigned mTICI scores were extracted from the clinical reports and reviewed by a neuroradiologist (>5 years experience) to correct for mismatches between extracted scores and exported

Table 1. Literature Values on Interrater and Intrarater Variability of TICI Scoring

Study	Design	N	Measure	Results
Bar et al ¹⁷	mTICI analogue rating by 2 experienced neuroradiologists	43	C _κ	0.39–0.45
Heiferman et al ¹⁹	Intra- and interrater reliability of mTICI ratings of 4 raters	67	C _κ	0.37 (interrater) 0.45–0.71 (intrarater)
Liebeskind et al ¹³	Dichotomic differentiation between mTICI 2b (50%–99% reperfusion) and TICI 2b (67%–99% reperfusion)	52	C _κ	0.82
Suh et al ¹⁸	mTICI and TICI score with additional 50% to 67% reperfusion class, 5 readers	73	ICC	0.73 (TICI) 0.67 (mTICI)
Tung et al ¹⁵	Comparison of TICI, mTICI, and eTICI, 2 readers	129	C _κ	0.58 (TICI) 0.62 (mTICI) 0.61 (eTICI)
Volny et al ¹⁶	eTICI, 3 individual ratings and 2 consensus ratings	89	K _κ	0.36–0.56 (individual) 0.82–0.86 (consensus)

Measures to assess the rating agreement: Cohen's kappa (C_κ), Krippendorff's alpha (K_κ), and the ICC. eTICI indicates expanded TICI; ICC, interclass correlation coefficient; mTICI, modified TICI; and TICI, Thrombolysis in Cerebral Infarction.

DSA series. Demographic and angiographic details for the dataset are given in Table 2.

The external dataset comprises 95 fully anonymized DSA series of 61 patients from the NEUROSQUAD study.³² Inclusion criteria were the same as for the in-house dataset. Patient information and information about the TICI scoring process were not available due to strict anonymization. To avoid issues due to differing scoring guidelines, the DSA series were rescored by 2 in-house neuroradiologists (>5 years experience; blinded to each other). If the raters disagreed, they reviewed the case and jointly selected a final score. The raters had access to the preintervention DSA data.

The retrospective study and the analysis of the anonymized data was approved by the institutional ethical review board of the medical association of the University of Hamburg (part of approval number WF-035/18).

DL: Network Architecture

Main technical challenges in dealing with DSA image series in DL architectures for TICI scoring are 2-fold. First, the DSA data are stored in 2 different views, that is, frontal and lateral orientation, which provide complementary information. Second, the image data are not only 2-dimensional but of

Table 2. Demographic and Clinical Information for the In-House Dataset

Item	Median (IQR) or percent (counts)	
	In-house dataset	External dataset
Demographics		
Age	74 (62–80)	NA
Female	54.6% (430/789 patients)	NA
Occlusion sites		
M1	47.0% (371/789 patients)	NA
ACI/carotid-T	23.4% (185/789 patients)	NA
Posterior circulation	17.5% (138/789 patients)	NA
M2	10.9% (86/789 patients)	NA
ACA	1.1% (9/789 patients)	NA
M1 occlusions: included patients	64.5% (236/371 patients)	NA
Overall included patients / included DSA series	29.9% (236/789 patients) with in total 491 DSA series	61 patients with 95 DSA series
mTICI score		
0	35.2% (173/491 DSA series)	29.5% (28/95 DSA series)
1	5.5% (27/491 DSA series)	2.1% (2/95 DSA series)
2a	15.1% (74/491 DSA series)	16.8% (16/95 DSA series)
2b	22.6% (111/491 DSA series)	27.4% (26/95 DSA series)
3	21.6% (106/491 DSA series)	24.2% (23/95 DSA series)

Further details on patient inclusion/exclusion are given in the [Data Supplement](#). For the external dataset, no demographic information is available because of data anonymization. DSA indicates digital subtraction angiography; IQR, interquartile range; and mTICI, modified Thrombolysis in Cerebral Infarction.

spatio-temporal nature. Thus, the image data dimensionality has to be reduced in an appropriate manner that maintains the relevant spatio-temporal image information about the perfusion process—but under the constraint that the DL system can still be efficiently trained. To exploit the spatio-temporal information of the DSA series, the developed DL system builds on gated recurrent units³³ (GRU; a specific type of recurrent neural networks) in combination with the EfficientNet architecture.³⁴ The latter is used to fuse and condense the information of corresponding frontal and lateral DSA image frames by a 2-arm image encoder with shared weights. To increase robustness (ie, to reduce undesired oscillations of the network output) and physiological plausibility of the DL training process, framewise pseudo-perfusion labels are generated that reflect the perceived monotonically increasing perfusion during initial contrast media inflow. To incorporate the uncertainties inherent to human observer-based TICI scoring and the ordinal character of the score, the system is trained using a hybrid loss function, consisting of a variant of the common binary cross entropy that incorporates label uncertainty and a variant of the mean squared error that drastically penalizes large differences of the predicted and the desired mTICI score. Details about the applied methods can be found in the underlying methodical conference contribution³⁵ and the [Data Supplement](#). Smaller changes that further improved the model performance compared with the conference contribution were the application of sigmoid functions for computation of the pseudo-perfusion labels (before: linear) and the use of stochastic weight averaging in combination with learning rate scheduling (before: no scheduler applied).

Experiments

To demonstrate the potential of the DL system and to assess its capabilities to generalize, the study comprised 3 experiments.

Experiment I was based on the in-house dataset and dedicated to the parameter optimization and initial evaluation of the DL system. The in-house dataset was maintaining the mTICI class frequencies of the entire dataset, randomly split into 401 DSA series used for training purposes and 90 DSA series used as validation dataset. Similar to literature studies on interobserver variability of TICI scoring, the mTICI classes 0 and 1 were considered as occluded states and, motivated by the vague definition of mTICI score 1 and related high interobserver variability, combined to mTICI score 0/1.¹⁵ To select the best performing model parameters and to avoid overfitting during network training, an early stopping strategy was employed. The model hyperparameters are summarized in the supplemental materials.

In experiment II, the model trained during experiment I, that is, using the in-house dataset, was directly applied to the external dataset. Thus, the performance of the trained model was evaluated using unseen data from a different center.

It is, however, well known that DL systems are often sensitive to details of the center-specific imaging protocols. In experiment III, the network weights were therefore updated by model retraining on the external data and corresponding mTICI scores, starting with the experiment I model weights. Because of the limited size of the external dataset, a 3-fold cross validation was performed (data randomly split into training, validation and test set per fold; original class frequencies were maintained).

Statistical Analyses

The evaluation was performed on the DSA series level. For each experiment, the agreement of the expert-assigned mTICI scores and the prediction of the DL system were assessed by the proportion of specific agreement (SA) and Cohen's kappa (C_k). Further, the accuracy (ACC) as well as a sliding window accuracy ($ACC_{\pm 1}$) of the network predictions compared with the expert scores were evaluated. For $ACC_{\pm 1}$, a mTICI score prediction directly adjacent to the expert-based mTICI score is, in addition to an exact match, also considered correct. $ACC_{\pm 1}$ therefore allows assessing the frequency of indisputably wrong network predictions: the lower the sliding window accuracy is, the more network predictions differ by at least 2 levels of the mTICI scale compared with the expert score. Finally, in line with related literature, SA, C_k and ACC of the dichotomized problem mTICI score $\leq 2a$ versus mTICI score $\geq 2b$ (open state) were evaluated.

RESULTS

The results for the 3 experiments are summarized in Table 3. SA, C_k and the accuracy values correspond to the evaluation datasets of the different experiments, that is, DSA data not seen by the model during training.

Experiment I was designed to illustrate the feasibility of automated mTICI scoring based on the in-house dataset. The evaluation of Cohen's kappa as one of the most common measures to assess the inter- and intraobserver agreement of TICI scoring in literature resulted in $C_k=0.61$ for the agreement of the network predictions and the expert mTICI scores (dichotomized classification: $C_k=0.91$). C_k values between 0.60 and 0.80 are usually interpreted to indicate substantial agreement. More important in the given context, the C_k value well matches the range of corresponding literature values on human expert TICI scoring agreement (Table 1). The class-specific SA values were between 63% and 84% (dichotomized classification: 96%), overall accuracy was 71% (dichotomized classification: 96%), and $ACC_{\pm 1}$ of 100% illustrates robustness and reliability of the DL system (no predictions far apart from the expert mTICI score).

Directly applying the model of experiment I to predict mTICI scores of the unseen external dataset from a different center (experiment II) resulted in C_k , SA, ACC, and $ACC_{\pm 1}$ values of 0.52, between 43% and 87%, of 64%, and of 97%, respectively (dichotomized problem: $C_k=0.73$, SA=86%/87%, ACC=86%). While C_k of 0.52 still indicates moderate agreement between the model predictions and the expert scores, the values represent a decrease of performance compared with experiment I and an observed $C_k=0.75$ for the external dataset ratings of the 2 in-house experts. In addition, the evaluation measures show considerable variations between the classes. For example, the model performs much better for class mTICI score 0/1 (ACC=90%, $C_k=0.81$) when compared with class 2b (ACC=39%, $C_k=0.24$). The drop of performance was, however, not unexpected due to

Table 3. Results of the Automated TICI Scoring Performance

Metric and experiment	mTICI score0/1	mTICI 2a	mTICI 2b	mTICI 3	Overall	Dichotomized
C_k exp. I	0.78	0.60	0.49	0.58	0.61	0.91
C_k exp. II	0.81	0.41	0.24	0.54	0.52	0.73
C_k exp. III	0.85	0.70	0.45	0.57	0.65	0.85
SA (%), exp I	84.0	68.4	62.5	68.2	NA	95.5/95.7
SA (%), exp II	87.1	51.4	42.6	65.2	NA	86.6/86.0
SA (%), exp III	89.7	75.7	60.4	66.7	NA	92.6/92.6
ACC (%), exp. I	84.0	65.0	60.0	75.0	71.1	95.6
ACC (%), exp. II	90.0	56.3	38.5	65.2	64.2	86.3
ACC (%), exp. III	86.7	87.5	61.5	60.9	73.7	92.6
$ACC_{\pm 1}$ (%), exp. I	100.0	100.0	100.0	100.0	100.0	NA
$ACC_{\pm 1}$ (%), exp. II	96.7	100.0	96.2	95.7	96.8	NA
$ACC_{\pm 1}$ (%), exp. III	96.7	100.0	96.2	100.0	97.9	NA

Cohen's kappa (C_k), the proportion of SA, and the accuracy measures (ACC, $ACC_{\pm 1}$) are computed based on the confusion matrices in the Figure and correspond to the in-house testing set (experiment I), the entire external dataset (experiment II), and the accumulated results of the cross validation performed on the external dataset (experiment III; corresponding results for the individual folds are given in the supplemental materials). The values are evaluated for the individual TICI classes, overall, and for the dichotomized classification mTICI score $\leq 2a$ vs mTICI score $\geq 2b$. mTICI indicates modified TICI; SA, specific agreement; and TICI, Thrombolysis in Cerebral Infarction.

different characteristics of the external image data (eg, late starting point of the DSA sequences) compared with the in-house data—and motivated experiment III.

In the 3-fold cross validation of experiment III, the model of experiment I was updated using 1/3 of the external dataset (ie, 31–32 DSA series) as training set for each fold. The model update led to an increase of the mTICI scoring performance measures; C_k , SA, ACC, and $ACC_{\pm 1}$ values of 0.65, between 60% and 90%, of 74%, and of 98%, respectively, are in the order of or higher than for experiment I and the in-house dataset and much closer to the agreement of the 2 experts for the external dataset. The measures for the dichotomized classification also increased and are in the order of corresponding values on interrater agreement.¹³

The performance gain/loss between experiments is visualized in the Figure, representing the confusion matrices for the 3 experiments.

Training and Inference Run-Time

The DL system was developed under the constraint of moderate run-time for model inference (the application of the model in a clinical setting) without dedicated DL hardware. The initial training of the network takes ≈ 90 minutes on a DL cluster. After training, the inference time was 0.7 seconds on a standard graphical processing unit and 3.7 seconds on the central processing unit of a standard working station for a DSA series with a spatial size of 1024×1024 pixels and a length of 15 frames. Details on the applied hardware are given in the [Data Supplement](#).

DISCUSSION

Addressing the need for objective TICI scoring in the context of MT treatment of AIS, we presented the first

(to the best of our knowledge, based on our recent conference contribution³⁵) approach to fully automated (m) TICI scoring, exploiting current DL trends for spatio-temporal image data. The emphasis of the performed experiments of this proof-of-principle study was to evaluate the general suitability of the DL model in terms of clinical applicability.

Focusing on M1 occlusions, the DL system achieved an agreement of network-predicted and expert-assigned mTICI scores in the range of published expert-based TICI scoring variability measures^{15–19} for 2 independent DSA datasets from different centers. In addition, the dichotomous distinction between open (mTICI score $\geq 2b$) and remaining states (mTICI score $\leq 2a$) was highly reliable. Respective performance measures are also in the range of corresponding literature values of interrater variability studies.^{13,19}

A direct application of the DL model trained on the in-house dataset to the external dataset revealed nevertheless a drop of the prediction accuracy for the different center data. Main reasons are most probably differences of the characteristics of the DSA image series of the 2 centers (start of series acquisition, applied amount of contrast agents, field of view, etc). This observation is common for DL and indicates an undesired overfitting of the model to these characteristics during training. In the present case, a computationally inexpensive model update using only a small number of training series allowed the network to adapt to the different center DSA series. After the update, the model performance for the external data was similar to the performance for the in-house data.

Focusing on the individual mTICI classes, the prediction performance was lowest for mTICI 2b for all experiments. The reason remains unclear. However, a review

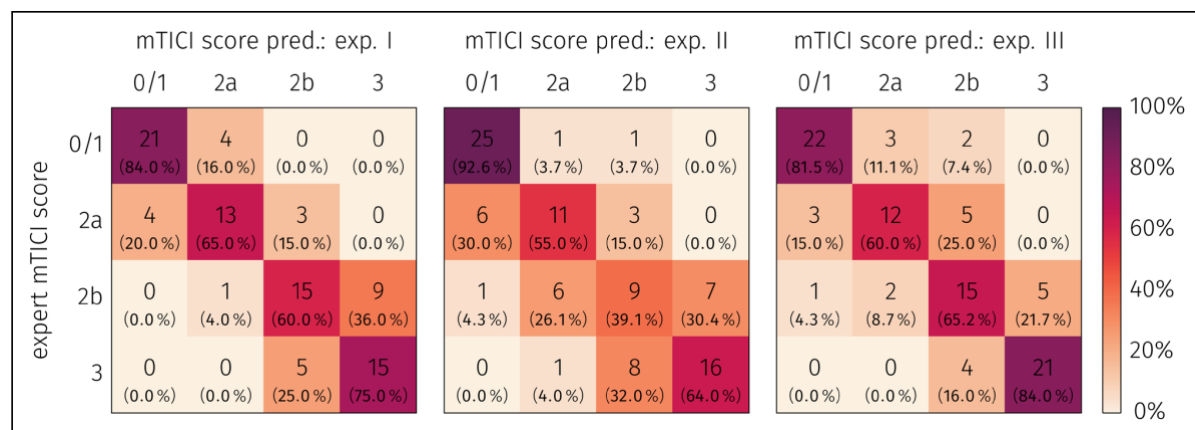


Figure. Confusion matrices for experiment I (left), II (middle), and III (right) with absolute values (top numbers in cells) and corresponding relative values (bottom numbers).

The absolute values in the confusion matrix for experiment III represent an accumulation of the values for the 3 cross validation folds. The confusion matrices for the individual folds are given in the [Data Supplement](#). The cell coloring scheme is based on the relative values.

of the mTICI 2b and 2a cases revealed that almost half of them show a reperfusion of the middle cerebral artery territory of $\approx 40\%$ to 60% , that is, close to the threshold of 50% that, by definition, differentiates between 2a and 2b. Interestingly, reported intrarater variability data suggest that this problem also exists for human readers^{19,36}—meaning that the medical expert-based scoring for cases close to mTICI-inherent reperfusion thresholds is not as reliable as ideally desired for training of neural networks. This, in turn, poses a fundamental problem and renders the interpretation of the accuracy values difficult. A possible solution to increase reliability and robustness might be a re-scoring of the DSA datasets by multiple clinical experts, that is, a consensus rating,¹⁶ instead of using single rater/clinical scores for initial model training. Consensus rating for large datasets is, however, very time-consuming and was beyond the scope of the present study. From the technical perspective, a workaround would be to analyze the mTICI class probabilities output by the network instead of a single mTICI score and/or to train and apply a network ensemble to estimate class-specific scoring probabilities and uncertainties. This would potentially allow identifying cases close to the class borders. Moreover, the classification could be reformulated as a continuous variable task to predict the actual degree of reperfusion after MT. However, in such a case, mTICI-inherent aspects like slow peripheral blood flow would not be covered.

From the application perspective, (1) the short run-time of 0.7 seconds (graphical processing unit) and 3.7 seconds (central processing unit) for standard DSA image series on standard computer hardware and (2) the capability of the proposed system to directly work on the clinical DSA image data (no expert preprocessing or curation of images necessary) indicate feasibility of direct integration into clinical workflows, for instance, to provide the attending physician with a score suggestion

with corresponding probabilities shortly after the intervention. This could be an important step toward more objective and consistent scoring within and between clinics. In a similar way, application of the DL model could reduce the reported mismatch between local and CoreLab scoring²⁰ as well as scoring variability within and between CoreLabs and thus lead to improved comparability of clinical studies.

With regard to an integration of DL-based automated TICI scoring into clinical workflows, the main limitations of the present proof-of-principle study are the strict patient selection criteria and the limited size of the resulting dataset (although larger than most available studies on TICI scoring^{13,15–19}). We expect extending the system to other occlusion sites than M1 as methodically straight-forward, but this remains to be shown and a corresponding dataset of sufficient sample size collected. Further, a deployed system should be capable of dealing with, for instance, additional pathologies in the images, foreign materials in the field of view, and the presence of image artifacts—or, at least, be able to identify and report respective issues. This, in turn, leads to the active field of explainable artificial intelligence and corresponding methodical developments. Finally, based on our observations for the external dataset, we expect, at the moment, model updating by means of exemplary cases beneficial for imaging data from centers that did not contribute to the training dataset. Alternative solutions could be the integration of approaches to data debiasing.

Focusing on the methodical contributions, the main advantages of the proposed end-to-end DL-based approach to automated mTICI scoring in comparison to semiautomatic methods that aim at the assessment of the antegrade reperfusion in brain tissue (and with that on an indirect TICI score assessment using perfusion values) like the recently described qTICI approach²³ are as follows:

- The mTICI score definition is learned implicitly. That is, the evidence of vessel revascularization, slow distal blood flow and antegrade perfusion are inherently taken into account.
- Training of the network is based on a single label, the expert mTICI score, per DSA series. This ensures sufficient data availability for training or updating the network, like for usage at additional centers, because such information is routinely collected during MT treatment anyway.
- The end-to-end DL architecture enables a multipurpose application of the system, like an extension to identify the occlusion site or a correlation of the network output with clinical outcome measures (eg, modified Rankin Scale score after 90 days or final NIHSS).

The latter aspect leads to next steps and follow-up studies: Although there are no arguments against it, it remains to be shown that the system can be trained for other occlusion sites than M1 and more detailed scores like expanded TICI. We further expect the system to benefit from additional training data, especially with regard to the mTICI 1 and 2a classes that were, due to their lower clinical frequency, underrepresented in our in-house dataset. Moreover, the model update in experiment III was based on only 31–32 DSA series per cross validation fold; it is to be shown whether this explains the difference between model-expert and expert-to-expert agreement observed for the external dataset. However, the performed experiments demonstrate that automated TICI scoring is feasible. Thus, the investigation of the system outputs with respect to clinical questions like correlation to patient outcome will be the next steps of our future work.

CONCLUSIONS

The present proof-of-principle study demonstrated feasibility of DL-based automated TICI scoring, making use of independent DSA datasets from 2 centers. For the considered patient collective and M1 occlusions, respectively, the agreement of the DL-based predictions and expert mTICI scores is in the range of published intra- and interrater variability and agreement measures of human expert-based TICI scoring. This highlights the future potential of the system to improve TICI score comparability between different raters, centers, and clinical studies, for example, by guiding physicians by means of model-predicted TICI scores.

ARTICLE INFORMATION

Received December 4, 2020; final revision received April 14, 2021; accepted April 22, 2021.

Affiliations

Department of Computational Neuroscience (M.N., T.S., R.W.), Center for Biomedical Artificial Intelligence (bAlome) (M.N., T.S., R.W.), and Department of Diagnostic and Interventional Neuroradiology (M.W., A.M.F., F.F., J.F.), University Medical Center-

Hamburg-Eppendorf, Germany; Eppdata GmbH, Hamburg, Germany (E.H.), Department of Neuroradiology, Heidelberg University Hospital, Germany (M.B., F.S.).

Acknowledgments

We thank Farid Kandil (Charité, Berlin, Germany) for providing support for the statistical data evaluation.

Sources of Funding

This work was supported by the European Fund for Regional Development and the Free and Hanseatic City of Hamburg.

Disclosures

Dr Flottmann reports unrelated personal fees from Eppdata. EH is with Eppdata. Dr Bendszus reports unrelated grants from DFG, EU, Siemens, Novartis, Guerbet, Hopp Foundation; and personal fees from Novartis, Guerbet, Codman, Teva, Bayer, Merck, BBraun, Grifols, Boehringer Ingelheim, Springer, Vascular Dynamics, Dr Fiehler reports related grants from CEO Eppdata, from Hamburgische Investitions- und Förderbank (IFB)/ERDF, unrelated grants from BMBF, BMWi, DFG, EU, grants/personal fees from Acandis, Cerenovus, Medtronic, Microvention, Route92, Stryker, personal fees from Bayer, Boehringer Ingelheim, Covidien, Evasec Neurovascular, Penumbra, Phenox, Transverse Medical; he is stock holder at Tegus Medical. Dr Werner reports related grant from IFB/ERDF and unrelated grants from Siemens Healthcare, DFG, Roggenbuck Stiftung. The other authors report no conflicts.

Supplemental Materials

Expanded Methods and Detailed Results
MI-CLAIM Checklist
Online Movies I–V

REFERENCES

1. Saver JL, Goyal M, Bonafe A, Diener HC, Levy EI, Pereira VM, Albers GW, Cognard C, Cohen DJ, Hacke W, et al; SWIFT PRIME Investigators. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med*. 2015;372:2285–2295. doi: 10.1056/NEJMoa1415061
2. Jovin TG, Chamorro A, Cobo E, de Miquel MA, Molina CA, Rovira A, San Román L, Serena J, Abilleira S, Ribó M, et al; REVASCAT Trial Investigators. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med*. 2015;372:2296–2306. doi: 10.1056/NEJMoa1503780
3. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, Roy D, Jovin TG, Willinsky RA, Sapkota BL, et al; ESCAPE Trial Investigators. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med*. 2015;372:1019–1030. doi: 10.1056/NEJMoa1414905
4. Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, Schonewille WJ, Vos JA, Nederkoorn PJ, Wermer MJ, et al; MR CLEAN Investigators. A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med*. 2015;372:11–20. doi: 10.1056/NEJMoa1411587
5. Turc G, Bhogal P, Fischer U, Khatri P, Lobotesis K, Mazighi M, Schellinger PD, Toni D, de Vries J, White P, et al; European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) guidelines on mechanical thrombectomy in acute ischemic stroke. *J Neurointerv Surg*. 2019;11:535–538. doi: 10.1136/neurintsurg-2018-014568
6. Ueda T, Hatakeyama T, Kohno K, Kumon Y, Sakaki S. Endovascular treatment for acute thrombotic occlusion of the middle cerebral artery: local intra-arterial thrombolysis combined with percutaneous transluminal angioplasty. *Neuroradiology*. 1997;39:99–104. doi: 10.1007/s002340050374
7. Ueda T, Sakaki S, Nochide I, Kumon Y, Kohno K, Ohta S. Angioplasty after intra-arterial thrombolysis for acute occlusion of intracranial arteries. *Stroke*. 1998;29:2568–2574. doi: 10.1161/01.str.29.12.2568
8. Ringer AJ, Qureshi AI, Fessler RD, Guterman LR, Hopkins LN. Angioplasty of intracranial occlusion resistant to thrombolysis in acute ischemic stroke. *Neurosurgery*. 2001;48:1282–1288. doi: 10.1097/00006123-200106000-00019
9. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, Yavagal DR, Ribo M, Cognard C, Hanel RA, et al; DAWN Trial Investigators. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med*. 2018;378:11–21. doi: 10.1056/NEJMoa1706442
10. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, McTaggart RA, Torbey MT, Kim-Tenser M, Leslie-Mazwi T, et al; DEFUSE 3 Investigators. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med*. 2018;378:708–718. doi: 10.1056/NEJMoa1713973

11. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, Biller J, Brown M, Demaerschalk BM, Hoh B, et al; American Heart Association Stroke Council. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2018;49:e46–e110. doi: 10.1161/STR.0000000000000158
12. Higashida RT, Furlan AJ, Roberts H, Tomsick T, Connors B, Barr J, Dillon W, Warach S, Broderick J, Tilley B, et al; Technology Assessment Committee of the American Society of Interventional and Therapeutic Neuroradiology; Technology Assessment Committee of the Society of Interventional Radiology. Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke. *Stroke*. 2003;34:e109–e137. doi: 10.1161/01.STR.0000082721.62796.09
13. Liebeskind DS, Bracard S, Guillemin F, Jahan R, Jovin TG, Majoie CB, Mitchell PJ, van der Lugt A, Menon BK, San Román L, et al; HERMES Collaborators. eTICI reperfusion: defining success in endovascular stroke therapy. *J Neurointerv Surg*. 2019;11:433–438. doi: 10.1136/neurintsurg-2018-014127
14. Zaidat OO, Yoo AJ, Khatri P, Tomsick TA, von Kummer R, Saver JL, Marks MP, Prabhakaran S, Kallmes DF, Fitzsimmons BF, et al; Cerebral Angiographic Revascularization Grading (CARG) Collaborators; STIR Revascularization working group; STIR Thrombolysis in Cerebral Infarction (TICI) Task Force. Recommendations on angiographic revascularization grading standards for acute ischemic stroke: a consensus statement. *Stroke*. 2013;44:2650–2663. doi: 10.1161/STROKEAHA.113.001972
15. Tung EL, McTaggart RA, Baird GL, Yaghi S, Hemendinger M, Dibiasio EL, Hidayat DT, Tung GA, Jayaraman MV. Rethinking thrombolysis in cerebral infarction 2b: which thrombolysis in cerebral infarction scales best define near complete recanalization in the modern thrombectomy era? *Stroke*. 2017;48:2488–2493. doi: 10.1161/STROKEAHA.117.017182
16. Volny O, Cimflova P, Szeder V. Inter-rater reliability for thrombolysis in cerebral infarction with TICI 2c category. *J Stroke Cerebrovasc Dis*. 2017;26:992–994. doi: 10.1016/j.jstrokecerebrovasdis.2016.11.008
17. Bar M, Mikulik R, Jonszta T, Krajina A, Roubec M, Skoloudik D, Prochazka V. Diagnosis of recanalization of the intracranial artery has poor inter-rater reliability. *AJNR Am J Neuroradiol*. 2012;33:972–974. doi: 10.3174/ajnr.A2896
18. Suh SH, Cloft HJ, Fugate JE, Rabinstein AA, Liebeskind DS, Kallmes DF. Clarifying differences among thrombolysis in cerebral infarction scale variants: is the artery half open or half closed? *Stroke*. 2013;44:1166–1168. doi: 10.1161/STROKEAHA.111.000399
19. Heiferman DM, Pecoraro NC, Wozniak AW, Ebersole KC, Jimenez LM, Reynolds MR, Ringer AJ, Serrone JC. Reliability of the modified TICI score among endovascular neurosurgeons. *AJNR Am J Neuroradiol*. 2020;41:1441–1446. doi: 10.3174/ajnr.A6696
20. Zhang G, Treurniet KM, Jansen IGH, Emmer BJ, van den Berg R, Marquering HA, Uyttenboogaart M, Jenniskens SFM, Roos YBWEM, van Doormaal PJ, et al; MR CLEAN Registry Investigators. Operator versus core lab adjudication of reperfusion after endovascular treatment of acute ischemic stroke. *Stroke*. 2018;49:2376–2382. doi: 10.1161/STROKEAHA.118.022031
21. Scalzo F, Liebeskind DS. Perfusion angiography in acute ischemic stroke. *Comput Math Methods Med*. 2016;2016:2478324. doi: 10.1155/2016/2478324
22. Shiraz Bhurwani MM, Snyder KV, Waqas M, Mokin M, Rava RA, Podgorsak AR, Chin F, Davies JM, Levy EI, Siddiqui AH, et al. Use of quantitative angiographic methods with a data-driven model to evaluate reperfusion status (mTICI) during thrombectomy. *Neuroradiology*. 2021;63:1429–1439. doi: 10.1007/s00234-020-02598-3
23. Prasetya H, Ramos LA, Epema T, Treurniet KM, Emmer BJ, van den Wijngaard IR, Zhang G, Kappelhof M, Berkhemer OA, Yoo AJ, et al; MR CLEAN Registry Investigators. qTICI: Quantitative assessment of brain tissue reperfusion on digital subtraction angiograms of acute ischemic stroke patients. *Int J Stroke*. 2021;16:207–216. doi: 10.1177/1747493020909632
24. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One*. 2014;9:e88225. doi: 10.1371/journal.pone.0088225
25. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke*. 2019;50:1263–1265. doi: 10.1161/STROKEAHA.118.024293
26. Nishio M, Koyasu S, Noguchi S, Kiguchi T, Nakatsu K, Akasaka T, Yamada H, Itoh K. Automatic detection of acute ischemic stroke using non-contrast computed tomography and two-stage deep learning model. *Comput Methods Programs Biomed*. 2020;196:105711. doi: 10.1016/j.cmpb.2020.105711
27. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, Ebert SA, Pomerantz SR, Romero JM, Kamalian S, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. 2019;3:173–182. doi: 10.1038/s41551-018-0324-9
28. Löber P, Stimpel B, Syben C, Maier A, Ditt H, Schramm P, Raczkowski B, Kemmling A. Automatic thrombus detection in non-enhanced computed tomography images in patients with acute ischemic stroke. The Eurographics Association, eds. In: VCBM'2017; 2017.
29. Kamal H, Lopez V, Sheth SA. Machine learning in acute ischemic stroke neuroimaging. *Front Neurol*. 2018;9:945. doi: 10.3389/fneur.2018.00945
30. Rai AT, Seldon AE, Boo S, Link PS, Domico JR, Tarabishy AR, Lucke-Wold N, Carpenter JS. A population-based incidence of acute large vessel occlusions and thrombectomy eligible patients indicates significant potential for growth of endovascular stroke therapy in the USA. *J Neurointerv Surg*. 2017;9:722–726. doi: 10.1136/neurintsurg-2016-012515
31. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaut R, Kohane IS, Saria S, Topol E, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320–1324. doi: 10.1038/s41591-020-1041-y
32. Seker F, Fiehler J, Möhlenbruch MA, Heimann F, Flottmann F, Ringleb PA, Thomalla G, Steiner T, Kraemer C, Brekenfeld C, et al. Time metrics to endovascular thrombectomy in 3 triage concepts: a prospective, observational study (NEUROSQUAD). *Stroke*. 2020;51:335–337. doi: 10.1161/STROKEAHA.119.027050
33. Chung J, Gulcehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. In: ICML'15. Proceedings of Machine Learning Research, vol. 37, pp. 2067–2075. JMLR.org; 2015.
34. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: ICML'2019. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. JMLR.org; 2019.
35. Nielsen M, Waldmann M, Sentker T, Frölich A, Fiehler J, Werner R. Time matters: handling spatio-temporal perfusion information for automated TICI scoring. Martel AL, et al, eds. In: MICCAI'2020, Lecture Notes in Computer Science, vol. 12266, pp. 86–96. Springer; 2020.
36. Mair G, von Kummer R, Adami A, White PM, Adams ME, Yan B, Demchuk AM, Farrall AJ, Sellar RJ, Ramaswamy R, et al; IST-3 Collaborative Group. Observer reliability of CT angiography in the assessment of acute ischaemic stroke: data from the Third International Stroke Trial. *Neuroradiology*. 2015;57:1–9. doi: 10.1007/s00234-014-1441-0

SUPPLEMENTAL MATERIAL

Deep learning-based automated TICI scoring: a timely proof-of-principle study

Maximilian Nielsen^{1,2,*}, MSc; Moritz Waldmann^{3,*}, PhD; Andreas M. Frölich³, MD;
Fabian Flottmann³, MD; Evelin Hristova⁴, MSc; Martin Bendszus⁵, MD; Fatih Seker⁵, MD;
Jens Fiehler^{3,‡}, MD; Thilo Sentker^{1,2,‡}, PhD; René Werner^{1,2,‡,†}, PhD

¹Department of Computational Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

²Center for Biomedical Artificial Intelligence (bAlome), University Medical Center Hamburg-Eppendorf, Hamburg, Germany

³Department of Diagnostic and Interventional Neuroradiology, University Medical Center-Hamburg-Eppendorf, Hamburg, Germany

⁴Eppdata GmbH, Hamburg, Germany

⁵Department of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany

* shared first authorship

‡ shared last authorship

† corresponding author

Supplemental A. Expanded Methods

Supplemental A.1. Network Architecture

In the following, modules of our workflow and network architecture, as schematically illustrated in Figure I, are briefly described. A more detailed explanation of the general methodological implementation can be found in our previous technical conference contribution (reference 35 of the original paper).

- *Pre-processing*

All frames of the 2D + t DSA series are prior to network training down-sampled to an image size of 224×224 pixels. Further, the intensity range is rescaled to an interval of $[-1, 1]$ after selecting relevant intensity thresholds by application of multi-Otsu thresholding. Standard data augmentation techniques (shot noise, blurring, shearing, flipping, cropping) are randomly applied.

- *Encoder*

The dimensionality of the pre-processed image data ($224 \times 224 \times t \times 2$) is further reduced by a two-arm image encoder, based on a pre-trained EfficientNet B0. The encoder fuses both views and densifies resulting data along the temporal axis, yielding an output feature size of 1280 per frame.

- *Gated recurrent unit*

The encoded image information is fed into a recurrent neural network that allows operating on frame-level information and information integration over time. Specifically, gated recurrent unit (GRU) networks are used, which, compared to commonly used long short-term memory (LSTM) networks, are more resource efficient at similar accuracy levels in, eg., sequence modeling and video representation learning.

For each time point of a given DSA series, the GRU yields a prediction. Contrast media (CM) inflow (visually perceived as perfusion) and retrograde perfusion is taken into account during the GRU learning process by utilizing frame-specific targets, which are computed by the implemented pseudo labeling process (see below).

- *Pseudo labeling*

The frame-specific output of the GRU is converted to a mTICI score and a corresponding perfusion value. Subsequently, the corresponding time points are normalized with respect to the DSA series length. Pseudo labeling now means to compute targets for these perfusion values and mTICI scores that match a priori assumptions. In our case, we hypothesize that the perceived perfusion should monotonically increase with an increasing normalized time and CM inflow until saturation. In turn, further increase of the perfusion after such a saturation would correspond to retrograde perfusion. The pseudo labels are then used to stabilize the learning process during temporal backpropagation. The proposed pseudo labeling process is based on

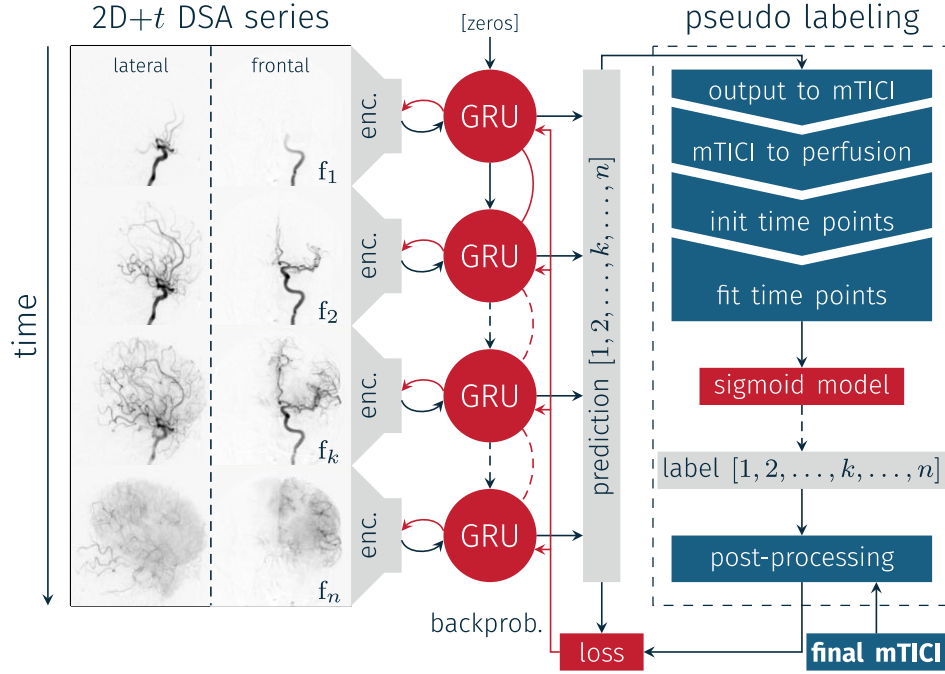


Figure I. Visualization of workflow and schematic network architecture. For each time point of the 2D + t DSA series, lateral and frontal views are encoded and combined by a two-arm encoder based on the EfficientNet B0 model. Subsequently, the encoded image information is fed into a gated recurrent unit (GRU), which yields a prediction for each time point of the DSA series. The output is converted to frame-specific perfusion values. Based on these values, frame-specific pseudo labels are computed by a simple sigmoid model, constrained by the actual clinical mTICI annotation. Finally, encoder and GRU are updated by back-propagating the loss between pseudo labels and predictions. During inference, the predicted mTICI score is determined by the model output for the last frame of the DSA series.

the assumption that CM-based perception of the perfusion process between first CM inflow and full patient-specific (re-)perfusion behind the occlusion site can be modeled by a simple model that meets the hypothesis, as shown in Figure II. In this study, the pseudo labels are computed by a sigmoid model constrained by the actual clinical mTICI annotation (ground truth). Resulting frame-specific labels are eventually post-processed, i.e., back-transformed into continuous mTICI targets and smoothed, before the loss is computed and back-propagated.

- *Loss function*

The loss is computed using a problem-specific uncertainty-driven TICI loss, implemented as a modified cross entropy combined with a mean squared error loss function to penalize larger deviations from the given clinical mTICI score. Here, the term *uncertainty-driven* can be interpreted as an uncertainty of the mTICI annotations with regard to the likelihood to assign a specific class to adjacent mTICI scores.

- *Network hyper-parameters and training*

Optimization is conducted using the Ranger optimizer with a learning rate of $5 \cdot 10^{-3}$ and cosine annealing to a minimum learning rate of $1 \cdot 10^{-4}$. A weight decay of $1 \cdot 10^{-3}$ is used. Stochastic weight averaging (SWA) is deployed after 85 epochs, where averaging is performed through an exponential running mean (more recent model weights are weighted higher) over 200 epochs. A loss uncertainty of 0.4 is applied. Dropout of 0.1 is employed in cell states of the GRU.

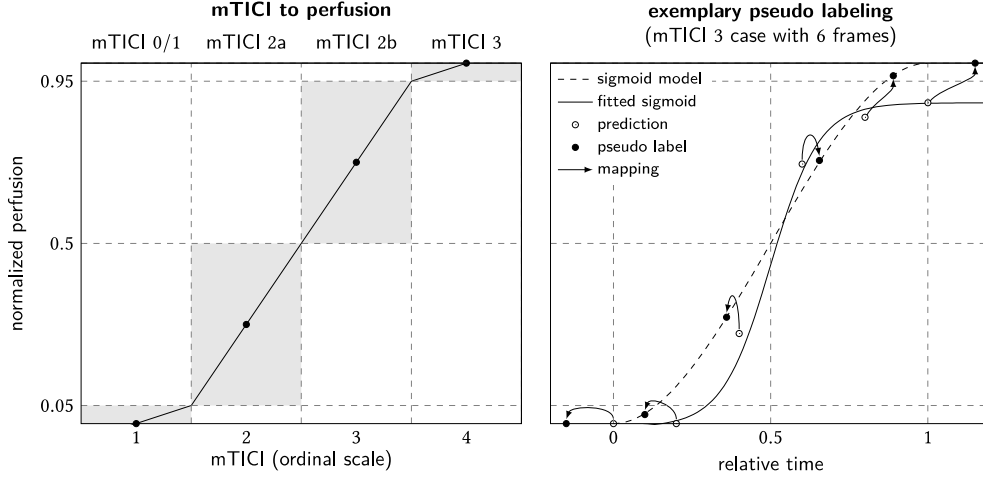


Figure II. Visualization of mTICI to perfusion conversion (left) and pseudo labeling (right). The exemplary pseudo labeling on the right shows the mapping between network-predicted perfusion values of a 6-frame DSA series with an expert-assigned mTICI score 3 (ie., perfusion finally equals 100%) to corresponding frame-specific pseudo ground truth perfusion values. Mapping between the values, visualized by arrows, is performed by fitting a sigmoid function (solid line) through the network predictions (empty circles) and shifting the fit to the corresponding pseudo ground truth sigmoid model (dashed line). Pseudo labels (filled circles) are yielded by applying resulting shift parameters to the network predictions. For further details see the underlying technical conference paper (reference 35 of the paper).

Network training is conducted on a GPU cluster equipped with four NVIDIA Tesla V100-32 GB GPU. With a batch size of 40, ie. 10 batches are processed per GPU per iteration (1 iteration equals $1/10$ epoch), the model converges after approximately 350 epochs, which corresponds to a training time of about 90 minutes.

• Model inference

During inference, the whole DSA series is processed and the predicted mTICI score is determined by the model output of the last frame. Inference time is approx. 0.5–1 s on standard GPUs and 3–4 s on CPU.

Supplemental A.2. Detailed Patient Selection (In-house Dataset)

A total of 789 patients with acute large-vessel occlusion ischemic stroke were referred to for the in-house data set. 418 of the initial 789 patients were excluded due to not matching the desired occlusion site, ie. occlusion of the M1 segment of the middle cerebral artery. The remaining 371 patients were further assessed for eligibility. Of the assessed patient datasets, 101 patients were excluded due to technical or image morphological reasons (acquisition problems such as less than five frames per DSA run, unsuitable representation of the respective territory, strong motion artifacts; see respective exclusion criteria in the main text), 18 patients did not receive a treatment (eg., caused by unsuccessful groin puncture or probing of the respective territory), for seven patients, treatment was performed using only one projection plane, five data sets contained (at least in one projection plane) intra- or extra-corporal foreign matter in the region of interest, and four patients showed additional pathology supra ophthalmic (dissection, SAB, aneurysm). The remaining 236 patient data sets with in total 491 DSA series were randomized and split in a train and test set of size 181 and 55, containing 401 and 90 DSA series, respectively; see also Fig. III.

Supplemental A.3. Statistical Analyses – Details of Multi-class C_K and SA Evaluation

In the present multi-class evaluation, Cohen's Kappa (C_K) and the proportion of specific agreement (SA) were computed. That is, consider the following exemplary $C \times C$ confusion matrix composed of elements a_{ij} , which denote the frequency of the cases that were scored as class i by a first rater (in our case the clinical expert) and class j by

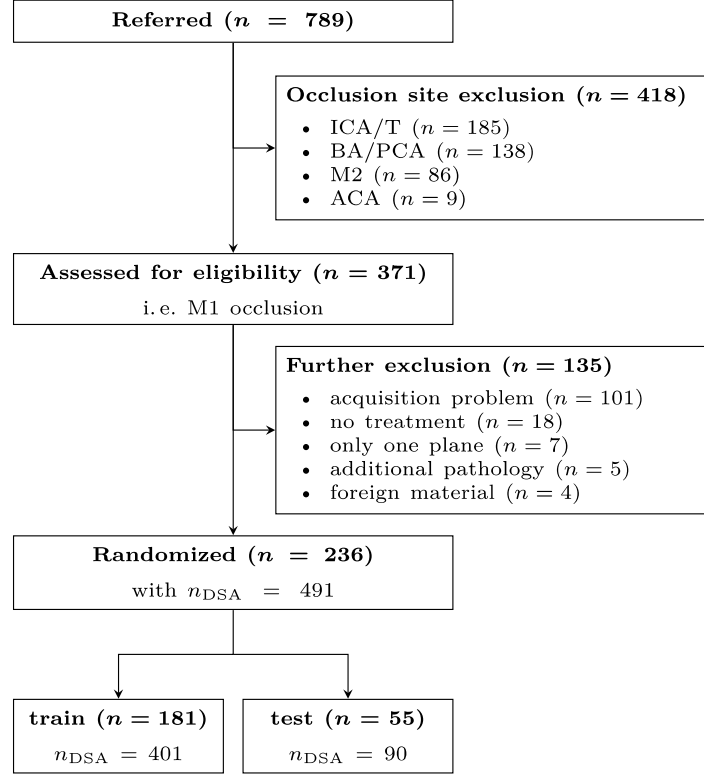


Figure III. Flow chart of patient in- and exclusion for the in-house dataset (ICA = internal cerebral artery, BA = basal artery, PCA = posterior cerebral artery, ACA = anterior cerebral artery).

a second rater (in our case the network). Table I illustrates such a matrix for C different classes where n_{*c} indicates the column sum, which corresponds to the amount of cases assigned to class c by the first rater. Accordingly, the row sum n_{c*} applies to rater two. The total number of cases is denoted by N .

Following this notation, Cohen's kappa that is defined as

$$C_\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (\text{A.1})$$

with p_o being the relative observed agreement and p_e the hypothetical probability of chance agreement, can be reformulated for a multi-class problem. The class-wise $p_{o,c}$ then reads:

$$p_{o,c} = \frac{N - n_{c*} - n_{*c} + 2a_{cc}}{N} \quad (\text{A.2})$$

and $p_{e,c}$:

$$p_{e,c} = \frac{n_{c*} \cdot n_{*c}}{N^2} + \frac{(N - n_{c*}) \cdot (N - n_{*c})}{N^2} \quad (\text{A.3})$$

Table I. Exemplary $C \times C$ confusion matrix.

		Model-predicted class				
		1	2	...	C	Σ
Expert-assigned class	1	a_{11}	a_{12}	...	a_{1C}	n_{1*}
	2	a_{21}	a_{22}	...	a_{2C}	n_{2*}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	C	a_{C1}	a_{C2}	...	a_{CC}	n_{C*}
	Σ	n_{*1}	n_{*2}	...	n_{*C}	N

For the overall (ie. consideration of all classes) evaluation, p_o is given by

$$p_{o, \text{overall}} = \frac{\sum_c a_{cc}}{N} \quad (\text{A.4})$$

and p_e reads

$$p_{e, \text{overall}} = \frac{1}{N^2} \sum_c (n_{c*} \cdot n_{*c}) \quad (\text{A.5})$$

SA is also calculated class-wise and for multi-class problems given by:

$$SA_c = \frac{2a_{cc}}{n_{c*} + n_{*c}} \quad (\text{A.6})$$

Supplemental B. Detailed Results

Figure IV shows the confusion matrices for the test sets for each of the three folds of the cross validation that was performed in experiment III. Corresponding numbers of the evaluated measures are summarized in Table II.

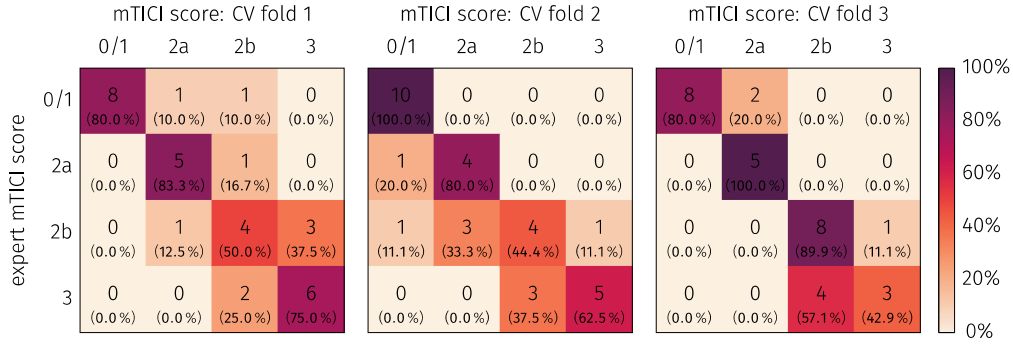


Figure IV. Confusion matrices for the test sets of the individual folds of the cross validation (CV) analysis in experiment III.

Table II. Results for each fold of the cross validation (CV) analysis of experiment III. The evaluation is based on the test set of the individual folds.

Metric	CV fold	mTICI 0/1	mTICI 2a	mTICI 2b	mTICI 3	Overall	Dichotomized
C_{κ}	1	0.85	0.71	0.33	0.60	0.62	0.81
	2	0.86	0.59	0.34	0.64	0.62	0.75
	3	0.84	0.79	0.64	0.46	0.69	1.00
SA (%)	1	88.9	76.9	50.0	70.6	n/a	90.3 / 90.9
	2	90.9	66.7	50.0	71.4	n/a	88.2 / 86.7
	3	88.9	83.3	76.2	54.5	n/a	100.0 / 100.0
ACC _{±1} (%)	1	90.0	100.0	100.0	100.0	96.9	n/a
	2	100.0	100.0	88.9	100.0	96.9	n/a
	3	100.0	100.0	100.0	100.0	100.0	n/a
ACC (%)	1	80.0	83.3	50.0	75.0	71.9	90.6
	2	100.0	80.0	44.4	62.5	71.9	87.5
	3	80.0	100.0	89.9	42.9	77.4	100.0

Supplemental C. Figures

Figure I summarizes the overall workflow of the presented DL system. Figure II sketches the conversion of the given ground truth mTICI scores for the considered DSA image series into a normalized perfusion value. Figure III illustrates the patient selection criteria applied to the in-house dataset. In Figure IV detailed results are visualized by confusion matrices for the test sets for each of the three folds of the cross validation that was performed in experiment III.

Supplemental D. Online Video Legends*Supplemental D.1. Examples for correct predicted cases of the in-house test set*

Video I (associated file: Nielsen-et-al_STROKE_0-0.mov): M1 occlusion with mTICI 0 score.

Video II (associated file: Nielsen-et-al_STROKE_2a-2a.mov): Original M1 occlusion with mTICI 2a score.

Video III (associated file: Nielsen-et-al_STROKE_3-3.mov): Original M1 occlusion with mTICI 3 score.

Supplemental D.2. Examples for incorrect predicted cases of the in-house test set

Video IV (associated file: Nielsen-et-al_STROKE_2a-2b.mov): Original M1 occlusion with mTICI 2a score (expert rating). The network predicts a mTICI 2b score.

Video V (associated file: Nielsen-et-al_STROKE_2b-3.mov): Original M1 occlusion with mTICI 2b score (expert rating). The network predicts a mTICI 3 score.

2 Einleitung

2.1 Behandlung des ischämischen Schlaganfalls

Die frühesten Beschreibungen eines Symptomkomplexes, welcher heute als Schlaganfall verstanden wird, datieren um 400 v. Chr. und werden Hippokrates von Kos zugeschrieben.¹ Hippokrates führte für das plötzliche Auftreten neurologischer Defizite den Oberbegriff Apoplex (vom griechischen *αποπληξια* - Niederschlag) ein, unter dem er auch weitere Krankheitsbilder wie Migräne und Epilepsien zusammenfasste und erklärte sie mit der damals verbreiteten Säftetheorie als Ansammlung von Galle oder Schleim in Hirnventrikeln.¹ Mitte des 17. Jahrhunderts erkannte der Schweizer Arzt und Anatom Johann Jakob Wepfer eine Blutung im Gehirn als möglichen Auslöser der Symptomatik und Anfang des 19. Jahrhunderts schlussfolgerte der schottische Arzt und Philosoph John Abercrombie, dass der Verschluss eines hirnzuführenden Gefäßes ursächlich für die Aufweichung des Gehirns ist, welche *post mortem* bei anderen Formen des Schlaganfalls beobachtet wurde.^{1,2} Rudolf Virchow prägte dann Anfang des 20. Jahrhunderts die noch heute gebräuchliche Unterscheidung in ischämischen und hämorrhagischen Schlaganfall.¹

Obwohl die wichtigsten Ursachen des Schlaganfalls erkannt waren, etablierte sich eine kausale, das heißt rekanalisierende Behandlung des ischämischen Schlaganfalls erst mit der *National Institute of Neurological Diseases and Stroke-2 Studie* (NINDS-2 Studie), welche 1995 im *New England Journal of Medicine* veröffentlicht wurde.³ Zuvor bestand die Standardtherapie ausschließlich in konservativen Maßnahmen wie einer Überwachung der Vitalparameter des Patienten, der Blutdruckeinstellung oder der Sauerstofftherapie sowie aus frührehabilitativen Maßnahmen. Die Ergebnisse der NINDS-2 Studie führten 1996 zur Zulassung des rekombinanten Gewebsplasminogenaktivators (*recombinant tissue plasminogen activator, rtPA*) zur Behandlung des ischämischen Schlaganfalls.⁴ Obwohl fibrinolytisch wirksame Enzyme bereits 1933 entdeckt und auch ab 1958 vereinzelt in der Schlaganfallbehandlung eingesetzt wurden, verzögerte sich die Etablierung dieser kausalen Therapieform initial auf Grund der noch nicht ubiquitär verfügbaren kranialen Bildgebung zur Differenzierung zwischen ischämischen und hämorrhagischen Schlaganfall und später durch das Fehlen einer therapeutischen Konsequenz, welche Anlass für eine Bildgebung gegeben hätte.⁴⁻⁷

Die begrenzte Wirksamkeit und das Blutungsrisiko der intravenösen Lysetherapie sowie die Erfahrungen bei der Behandlung des Myokardinfarkts führte zur Entwicklung der mechanischen Thrombektomie bei der Behandlung des ischämischen Schlaganfalls. Die lokale Therapie durch intraarterielle Gabe von Streptokinase Anfang der 80er Jahre durch Hermann Zeumer markiert den Beginn dieser Technik, welche sich in den folgenden 20 Jahren über die Angioplastie und das Stenting bis zur mechanischen Thrombektomie weiterentwickelt hat.⁸⁻¹² 2004 wurde der Merci Retriever X4 als erstes mechanisches Thrombektomieinstrument von der FDA zur Bergung zerebraler Emboli zugelassen und mit der Zulassung des Penumbra Systems 2007 mit der Aspiration thrombotischen Materials um eine Facette ergänzt.¹³ Der Durchbruch der mechanischen Thrombektomie erfolgte 2015 in Folge der Veröffentlichung von fünf großen, randomisierten Studien (MR CLEAN, ESCAPE, SWIFT PRIME, REVASCAT, EXTEND-IA), welche an insgesamt 1218 Patienten die mechanische Thrombektomie, mit oder ohne vorangehende intravenöser Lysetherapie, mit der konventionellen Therapie, bestehend aus Überwachung auf einer Stroke Unit und intravenöser Lysetherapie, verglichen.¹⁴⁻¹⁸ Die Studien zeigten einheitlich eine höhere Funktionalität und weniger Residuen bei den Patienten der Interventionsgruppe bei vergleichbarer Blutungs- und Sterblichkeitsrate.¹⁴⁻¹⁸ In der Folge wurde die mechanische Thrombektomie in diverse nationale Versorgungsleitlinien aufgenommen, die Technik weiterentwickelt und die Indikation erweitert, so dass die mechanische Thrombektomie heute ein etabliertes Verfahren der Behandlung des ischämischen Schlaganfalls ist.^{13,19-23}

2.2 Der TICI Score

Der TICI Score (*thrombolysis in cerebral infarction score*) wurde 2003 von Randall Higashida und Anthony Furlan unter der Schirmherrschaft der amerikanischen Gesellschaft für interventionelle und therapeutische Neuroradiologie sowie der amerikanischen Gesellschaft für interventionelle Radiologie als Teil einer Vereinheitlichung des Studiendesigns und der Dokumentation des Behandlungserfolges bei der Therapie des ischämischen Schlaganfalls vorgeschlagen.²⁴ Dies geschah vor dem Hintergrund, dass in klinischen Studien zur Therapie des Schlaganfalls unterschiedliche Kriterien und Definitionen des Behandlungserfolges verwendet wurden, was die Vergleichbarkeit zwischen den Studien erschwerte und damit einer Weiterentwicklung der Therapie im Wege stand.²⁴ Die ursprüngliche Definition war dabei an dem in der Kardiologie seit den 80ern gebräuchlichen TIMI *grade flow (thrombolysis*

in myocardial infarction score) angelehnt und wurde in der Arbeit von 2003 wie folgt in Tabelle 1 definiert²⁴:

Tabelle 1: Ursprüngliche Definition des TICI Scores nach Higashida et al. von 2003 in etwas gekürzter Wiedergabe.²⁴

TICI 0	Kein antegrader Fluss distal der Okklusion
TICI 1	Minimaler antegrader Fluss distal der Okklusion
TICI 2	Partielle Perfusion distal der Okklusion mit verzögertem Kontrastmittel Zu- und/oder Abfluss
TICI 2a	Kontrastmittelfüllung des distal zur Okklusion gelegenen Gefäßbettes zu weniger als $\frac{2}{3}$
TICI 2b	Vollständige, jedoch verzögerte Kontrastmittelfüllung des distal der Okklusion gelegenen Gefäßbettes
TICI 3	Reguläre und vollständige Kontrastmittelfüllung des Gefäßbettes

Der auf angiographische Bilddaten beruhende TICI Score wurde initial zur Dokumentation des Therapieerfolges der intraarteriellen Lysetherapie vorgeschlagen und fand schnell breite Akzeptanz, so dass er auch bald zur Beschreibung des Behandlungserfolgs bei mechanischer Rekanalisation herangezogen wurde. Die vage Definition des TICI Scores durch Higashida und Furlan führte in der Folge jedoch immer wieder zu unterschiedlichen Auslegungen und die fragliche klinische Relevanz der willkürlich definierten Grade zur wiederholten Überarbeitungen des TICI Scores mit teils unvollständiger Penetranz der Neuerungen.²⁵⁻²⁷ Teile der Ergebnisse einer Studie von 2013, welche die uneinheitliche und unvollständige Verwendung des TICI Scores untersuchte, sind in Abbildung 1 wiedergegeben.²⁸ So gibt es zum Beispiel der ursprünglichen Definition folgend keinen Score für die zwar nicht vollständige, dennoch aber über $\frac{2}{3}$ liegende Reperfusion eines zuvor okkludierten Gefäßterritoriums und auch keine Definition, wie ein interventionell nach distal translozierter Thrombus bei fortbestehender, vollständiger Okklusion zu bewerten ist.

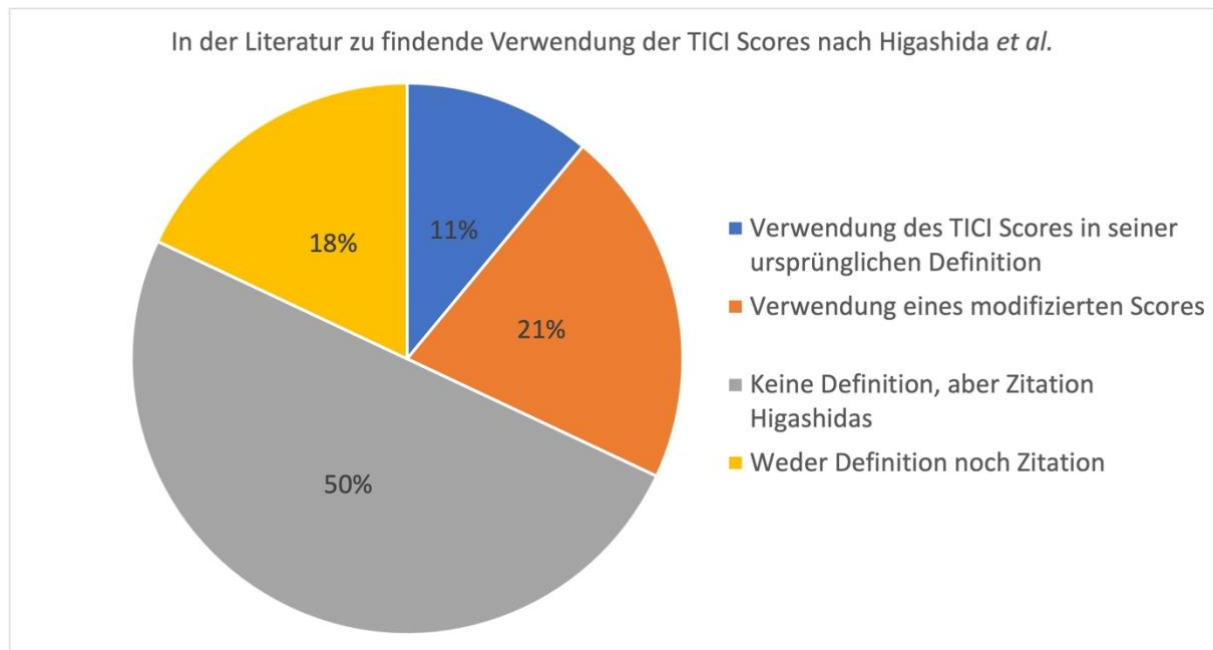


Abbildung 1: Kreisdiagramm mit Darstellung der in der Literatur zu findenden Verwendung des TICI Scores bezogen auf seine ursprüngliche Definition nach Higashida *et al.* in Prozent. Ausgewertet wurden 74 Studien aus 2004 bis 2012, die sich auf den TICI Score beziehen. In der Hälfte der Studien wurden die Originalarbeit ohne weitere Definition zitiert (gelb), in über 20% der Studien wurde der TICI abgewandelt (orange) und in nur etwas über 10% der Studien wie original beschrieben verwendet (blau).²⁸

Feinere Abstufungen des TICI Scores zwecks besserer prognostischer Vorhersage bezüglich des klinischen *outcome* wie TICI 2c, welcher eine nahezu vollständige Rekanalisation oder eine nur verzögerte Reperfusion des zuvor verschlossenen Gefäßbettes beschreibt, wurden vorgeschlagen und fanden weitere Anwendung in der Literatur.^{29,30} Zur Vereinheitlichung wurde 2013 in einem Konsensuspapier der mTICI Score (*modified TICI*) als Standardscore vorgeschlagen, welcher die Grenze zwischen TICI 2a und TICI 2b zu Gunsten einer höheren Interrater-Reliabilität und klinischen Relevanz auf 50% Füllung des Gefäßterritoriums festsetzt (Tabelle 2).²⁶

Tabelle 2: Definition des 2013 in einem Konsensuspapier vorgeschlagenen mTICI Scores.²⁶

mTICI 0	Keine Perfusion distal der Okklusion
mTICI 1	Geringe oder verzögerte antegrade Reperfusion distal der initialen Okklusion
mTICI 2a	Antegrade Reperfusion von bis zu 50% des zuvor verschlossenen Gefäßbettes
mTICI 2b	Antegrade Reperfusion von mehr als 50% des zuvor verschlossenen Gefäßbettes
mTICI 3	Vollständige Reperfusion des zuvor verschlossenen Gefäßbettes

2019 wurde der eTICI (*extended* TICI) Score vorgeschlagen, der zwecks weiterer Annäherung an das klinische *outcome* den mTICI 2b weiter unterteilt und zusätzlich die Kategorie TICI 2c einführt, womit der Score nun insgesamt sieben Graduierungen umfasst.²⁷ Die uneinheitliche Verwendung sowie die mehrfachen Überarbeitungen des TICI Scores zeigen aber auch die Probleme und Unzulänglichkeiten bei seiner Verwendung. Zwar ist er zur bildmorphologischen Beschreibung des Behandlungserfolges bei Rekanalisation sehr gut geeignet, jedoch korreliert er nur begrenzt mit dem klinischen *outcome* und unterliegt einer gewissen Interrater-Reliabilität.²⁷ Ein weiterer Nachteil ist, dass der TICI Score rein bildmorphologisch uneindeutig ist, das heißt, dass zu seiner Bestimmung die initiale Verschlusslokalisation bekannt sein muss. So entspricht das Bild einer perfundierten A. cerebri anterior bei verschlossener A. cerebri media einem TICI 0 bei initialem M1-Verschluss, jedoch einem TICI 2a bei initialem Verschluss der A. carotis interna.

2.3 Untersuchung zur Interrater-Reliabilität bei der Bestimmung des TICI Scores

Trotz der heute flächendeckenden Verwendung des TICI Scores ist die Studienlage bezüglich der Interrater-Reliabilität gering und die veröffentlichten Ergebnisse sehr unterschiedlich. Eine 2018 publizierte Studie untersuchte retrospektiv den Zusammenhang des eTICI Scores mit dem Patientenoutcome anhand der Daten der HERMES Kollaboration, welche die Patientendatensätze fünf großer Studien (MR CLEAN, ESCAPE, REVASCAT, SWIFT PRIME und EXTEND IA) einschließt. Dabei wurde unter anderem beobachtet, dass der vergebene Score nicht nur von der Erfahrung des *raters*, sondern auch von der Nähe zur durchführenden Institution abhängt. Externen Einrichtungen, wie sie zum Beispiel im Rahmen von Studien zur Quantifizierung des Behandlungserfolges hinzugezogen werden, sind in der Regel zurückhaltender bei der Bewertung bezüglich eines wünschenswerten Behandlungserfolges als es die Behandler vor Ort sind.²⁷ Es wurde weiter exemplarisch an 52 Patientendatensätzen

die dichotome Unterscheidung zwischen mTICI (*modified* TICI) TICI 2b Grenze 50% Reperfusion) und oTICI (*original* TICI) TICI 2b Grenze 67% Reperfusion) durch zwei *rater* untersucht, wobei als statistischer Güteparameter ein Cohens κ von 0.82 ($p < 0.001$) bestimmt wurde, was als sehr gute Übereinstimmung zwischen diesen beiden *ratern* interpretiert werden kann.²⁷

Eine 2017 publizierte Studie untersuchte retrospektiv die Interrater-Reliabilität des eTICI Scores an 89 Patientendaten in einem komplexen Schema, bei dem individuell erhobene Scores durch je einen angehenden neuroradiologischen Interventionalisten, der nicht an der Thrombektomie beteiligt war, einen Neuroradiologen und einen Schlaganfall-Neurologen verglichen wurden mit der Konsensus-Meinung des durchführenden Interventionalisten und dem anschließend behandelndem Arzt sowie weiter mit der Konsensus-Meinung aus Neuroradiologen und Schlaganfall-Neurologen.³¹ Das gemeinsam getroffene Scoring von durchführendem Interventionalisten und behandelndem Arzt wurde als Goldstandard herangezogen und als statistischer Güteparameter Krippendorfs α verwendet.³¹ Dabei wurden für die individuell bestimmten Scorings schlechte bis mäßige Übereinstimmungen (α zwischen 0.36 – 0.56) ermittelt und erst eine sehr gute Übereinstimmung bei Vergleich mit der Konsensus-Meinung von Neuroradiologen und Neurologen ($\alpha = 0.82$) erhalten.³¹ Für eine monozentrische Studie mit einer anzunehmend einheitlichen Auslegung des TICI Scores ist die Übereinstimmung trotz der detaillierteren Graduierung des eTICIs in insgesamt sechs Kategorien insbesondere für die individuell bestimmten Scores unbefriedigend.

Eine 2013 erschienene Studie untersuchte die Interrater-Reliabilität zwischen oTICI und mTICI, wobei eine Zwischenkategorie für eine Reperfusion zwischen 50 und 67% eingeführt wurde.³² Für beide Scores wurden befriedigende Werte von 0.73 bzw. 0.67 für den Interklassenkorrelationskoeffizient als statistischen Güteparameter bei der Bewertung von 73 Angiographieserien durch fünf unabhängige Rater erhalten.³² Darüber hinaus wurde beobachtet, dass etwa 20% der Patienten in die Zwischenkategorie zwischen 50% und 67% fallen, welche sich naturgemäß bei der Bewertung schwieriger darstellt.³²

In einer anderen Studie wurde ebenfalls über eine schlechte Übereinstimmung selbst zwischen zwei erfahrenen Neuroradiologen berichtet, welche retrospektiv und unabhängig voneinander Angiographieserien mechanischer Thrombektomien analog zum mTICI bewerteten.³³ Zur Auswertung wurde ein gewichtetes Cohens κ verwendet, bei dem die Ordinalskalierung des TICI Scores berücksichtigt wird und zum Beispiel die Unstimmigkeit

zwischen zwei benachbarten Klassen anders bewertet wird als eine größere Meinungsverschiedenheit. Für die Bewertung von 43 Patientenserien wurde das gewichtete Cohens κ auf 0.39 für Mediainfarkte und auf 0.45 für nicht Mediainfarkte bestimmt.³³ Die Autoren weisen explizit auf die Konsequenzen wie eine problematische Vergleichbarkeit klinischer Studien hin, welche sich aus einer so schlechten Übereinstimmung des Behandlungserfolges selbst unter erfahrenen Neuroradiologen ergeben.³³ Die Ergebnisse der verschiedenen Studien sind in Tabelle 3 zusammengefasst.

Tabelle 3: Zusammenfassung der in der Literatur beschriebene Ergebnisse zur Interrater-Reliabilität des TICI Scores. Die unterschiedlichen Herangehensweisen und Fragestellungen der einzelnen Studien sind unter Design wiedergegeben. Die jeweils verwendeten statistischen Güteparameter bei den Ergebnissen benannt. Vereinfacht lässt sich hierzu sagen, dass die für diese Fragestellung verwendeten Statistikparameter den Zahlenraum zwischen 0 und 1 umfassen, wobei eine Annäherung an 1 einer höheren Übereinstimmung entspricht. Auffallend sind weiter die relativ kleinen Fallzahlen, an denen die Interrater-Reliabilitäten untersucht wurden.

Studie	Liebeskind et al. 2018 ²⁷	Volny et al. 2016 ³¹	Suh et al. 2013 ³²	Bar et al. 2012 ³³
Design	Unterscheidung zwischen mTICI 2b (Grenze 50% Reperfusion) und oTICI 2b (Grenze 67% Reperfusion)	Individuell bestimmtes und konsensbasiertes Scoring nach eTICI	m- und oTICI Score mit extra Klasse für Reperfusion zwischen 50 – 67% (5 rater)	mTICI analoges Scoring durch zwei erfahrene Neuroradiologen
n Pat.	52	89	73	43
Basiert	Multizentrische erhobene Daten, monozentrisches Scoring	Monozentrische Studie	Monozentrische Studie	Monozentrische Studie
Ergebnisse	Cohens κ 0.82	Krippendorffs α zwischen 0.36 – 0.56 für individuell bestimmte Scores und 0.82 für Konsensus Meinungen	Interklassenkorrelationskoeffizient 0.73 bzw. 0.67	Gewichtetes Cohens κ 0.39 bzw. 0.45

Eine Studie von 2018 verglich an 1130 Patientendaten aus der MR CLEAN Studie den vor Ort durch den Interventionalisten vergebenen mTICI Score mit dem durch eine zentrale Institution vergebenen Score, wobei nur in 56% der Fälle eine Übereinstimmung beobachtet wurde.³⁴ Diese an einer großen Fallzahl erhobene Studie unterstreicht die Schwierigkeit bei der

Vergabe des TIC1 Scores und zeigt die Wichtigkeit eines institutsübergreifenden objektiven Quantifizierung des Behandlungserfolgs beim ischämischen Schlaganfall.

2.4 *Deep learning* Ansätze in der medizinischen Bildverarbeitung

Deep learning beschreibt eine Form von künstlicher Intelligenz, bei der neuronale Netzwerke selbstständig erlernen, Probleme zu lösen, für die es klassischerweise menschlicher Intelligenz bedarf.³⁵ Neuronale Netze bestehen aus digitalen Knotenpunkten angeordnet in einer Mehrschichtarchitektur aus einem *input*, einem *output* sowie einer variablen Zahl an *hidden layers*, deren Verschaltung entfernt an die von Neuronen erinnert (Abbildung 2).³⁵ Um bei gegebenem *input* zum gewünschten *output* zu gelangen, werden diese Netze mit Hilfe großer Datenpools trainiert und mittels mathematischer Algorithmen wie der *backpropagation* optimiert.³⁶ Das „*deep*“ in *deep learning* bezieht sich dabei auf die Anzahl der *hidden layer*, wobei dafür keine klare Grenze definiert ist.³⁶

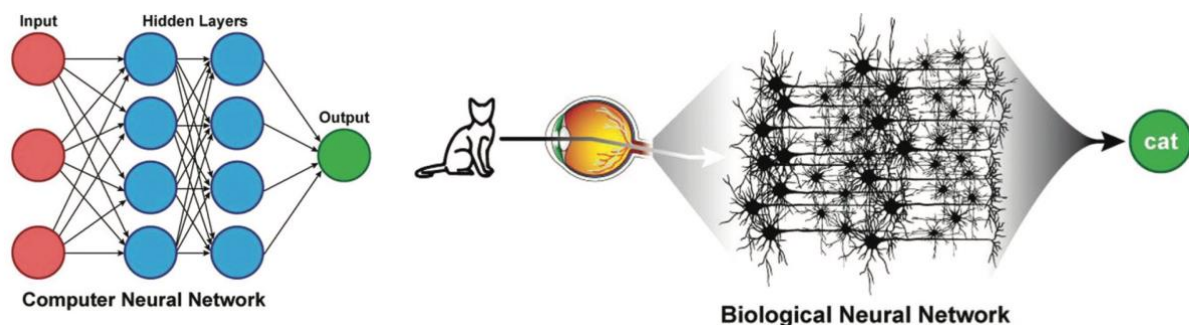


Abbildung 2³⁵: Anordnung und Verschaltung digitaler Knotenpunkte, sogenannter Neurone, auf unterschiedlichen Ebenen in einem neuronalen Netzwerk (links). Der Aufbau bestehend aus einem input layer, einem output layer sowie einer variablen Zahl an hidden layers. Dieser Aufbau erinnert an die Verschaltung von Neuronen in einem biologischen System (rechts), exemplarisch gezeigt an der visuellen Verschaltung bei der Interpretation (output) eines gegebenen inputs (Katze).

Vor gut fünf Jahren erlebten *deep learning* Algorithmen in der Beforschung zu medizinischen Applikationen ihren Durchbruch. Eine pubmed Suche nach „*deep learning*“ zeigt für Jahrzehnte konstante Publikationszahlen im niedrig zweistelligen Bereich, bis sich seit 2015 die Zahl der veröffentlichten Studien dazu jedes Jahr nahezu verdoppelten und für 2020 über 9000 Veröffentlichungen umfasst. Dieser rasant steigenden Zahl von Veröffentlichungen war ein Erdrutschsieg von neuronalen Netzen mit spezieller Architektur, sogenannten Konvolutionsnetzen (engl. *convolutional neural networks*, CNN), bei einem international anerkannten Wettbewerb der KI-Szene, dem ImageNet Large Scale Visual Recognition Challenge (ILSVRC), zur automatisierten Bilderkennung 2012 voraus gegangen.³⁷ Durch ihren speziellen Aufbau erreichen Konvolutionsnetze eine räumliche Invarianz bestimmter

Eigenschaften bei gleichzeitig reduziertem Rechenaufwand, was sie sehr robust bezüglich Position und Rotation des zu erkennenden *features* macht, wodurch sie sich besonders für eine Applikation auf mehrdimensionale Daten wie Bilddaten eignen.³⁸ Konvolutionsnetze zeigten dabei in Studien bei der Beantwortung definierter Fragestellungen bezüglich medizinischer Bilddaten wie Funduskopiebildern, dermatoskopischen Aufnahmen oder histopathologischen Schnitten Leistungen auf menschlichem Niveau.³⁹⁻⁴¹ Auch für radiologische Anwendungen wird die Implementierung von *deep learning* Algorithmen intensiv beforscht, wobei der Schwerpunkt auf den Bereichen Detektion, Segmentierung, Prozessierung und Klassifikation liegt.^{42,43} Getrieben wird das Interesse von dem Wunsch nach höherer Effizienz bei einer gleichzeitig erniedrigter Fehlerrate, da sich speziell in der Radiologie in den letzten 20 Jahren ein zunehmendes Missverhältnis von akquirierten Bilddaten zu qualifizierten Befundern entwickelt hat.⁴⁴

Ein weiterer Vorteil von der Applikation künstlicher Intelligenz liegt in der intra- und inter-institutionellen Vergleichbarkeit ihrer Ergebnisse, da sie, einmal trainiert, ohne menschlichen Beitrag agieren, was neben der allgemeinen Zeitersparnis bei ihrer Verwendung speziell für Studien und im Bereich des Qualitätsmanagements von Bedeutung ist.

2.5 Training neuronaler Netzwerke

Um bei gegebenem *input* zum gewünschten *output* zu gelangen, müssen neuronale Netzwerke trainiert werden. Bei dem Training werden mit Hilfe von Datensätzen die Verbindungen der Neuronen zwischen den einzelnen Schichten in iterativen Prozessen unterschiedlich gewichtet und so die Wahrscheinlichkeit, zum gewünschten Ergebnis zu gelangen, maximiert.³⁶

Das heute am meisten verwendete Verfahren ist das *supervised learning*, bei dem das Training mit Hilfe gelabelter Daten erfolgt. Das Label dient dabei als Grundwahrheit (engl. *ground truth*), an Hand welcher die Optimierung der Gewichtung der verschiedenen Verknüpfungen erfolgt.³⁶ Die Abweichung des erhaltenen vom gewünschten Ergebnis wird mit einer Kostenfunktion quantifiziert, wobei das Ergebnis die Güte des neuronalen Netzwerks für die gestellte Aufgabe beschreibt. Durch verschiedene mathematische Operationen wie der Rückwärtspropagierung (engl. *backpropagation*) werden die Verknüpfungen zwischen den einzelnen Neuronen für jeden einzelnen Trainingsfall - Neuron für Neuron, Schicht für Schicht

rückwärts durch das Netzwerk schreitend - so adjustiert, dass das Ergebnis der Kostenfunktion minimiert wird.³⁸

Eine wesentliche Limitation bei dem Training neuronaler Netzwerke für die unterschiedlichsten Applikationen stellt die Verfügbarkeit gut gelabelter Datensätze dar. Die Leistungsfähigkeit des Netzwerks hängt neben der Komplexität der gestellten Aufgabe zum einen von der Qualität der Labels und zum anderen von der Größe des Datensatzes ab, wobei speziell Konvolutionsnetze als sehr datenhungrig gelten.³⁸

3 Zielsetzung

Bei der Behandlung des ischämischen Schlaganfalls, welcher als dritthäufigste Todesursache und häufigste Ursache für erworbene Behinderungen in Deutschland von hoher epidemiologischer Relevanz ist, gewinnen mechanische Rekanalisierungsverfahren in den letzten Jahren zunehmend an Bedeutung.^{15,20,45,46} Der Erfolg dieser Therapie ist verglichen mit dem etablierten Standardverfahren, der intravenösen Lysetherapie, stärker von dem behandelnden Arzt und der therapiebezogenen Expertise am durchführenden Zentrum abhängig.^{47,48} Der bildmorphologische Erfolg der mechanischen Thrombektomie (MT) wird mittels des TICI Scores quantifiziert, wobei die Zuordnung anhand der während der Therapie generierten angiographischen Bilddaten visuell durch den behandelnden Arzt erfolgt.²⁶ Diese individuell variable Einteilung und die damit einhergehende begrenzte Vergleichbarkeit des Behandlungserfolges ist sowohl für das Qualitätsmanagement, zum Beispiel im Rahmen von Zertifizierungen zur Durchführung des Therapieverfahrens, als auch im Rahmen klinischer Studien problematisch.

In Zusammenarbeit mit dem *Institute for Computational Neuroscience* sollte eine Machbarkeitsstudie zur automatisierten Beurteilung des Therapieerfolges der MT durch neuronale Netzwerke durchgeführt werden. Speziell sogenannte Konvolutionsnetze (CNN) haben in den letzten Jahren eine rasante Entwicklung erfahren und sich als ausgesprochen leistungsfähige Algorithmen bei der Beurteilung und Verarbeitung von Bilddaten etabliert. Zur erfolgreichen Durchführung der gestellten Aufgabe sollten diese mit Hilfe eines Trainingsdatensatzes trainiert werden, wobei Größe und Qualität des Trainingsdatensatzes entscheidend für die Prädiktionsgenauigkeit des Netzwerks sind.³⁶

Im Rahmen der vorliegenden Arbeit sollte die Hypothese geprüft werden, ob neuronale Netzwerke in der Lage sind, angiographische Bildserien von Thrombektomien auf menschlichem Niveau nach TICI zu scoren (Hypothese 1). Dazu sollte ein Datensatz retrospektiv mit Hilfe einer klinikinternen Datenbank erhoben und neuronale Netzwerke mit diesem trainiert werden. Die Robustheit des Netzwerks bezüglich des Scorings sollte nach dem Training auf einen externen Datensatz untersucht und die Vorhersagegenauigkeit mit in der Literatur beschriebenen Werten zur Interrater-Reliabilität verglichen werden. Da es sich hierbei um eine Machbarkeitsstudie handelt, wurden enge Inklusionskriterien zur Aufnahme an den Datensatz gestellt. So sollten zum Beispiel nur Verschlüsse des vorderen Kreislaufs Berücksichtigung finden.

Der TICI Score beruht auf der Quantifizierung der Reperfusion bezogen auf die initiale Verschlusslokalisation und ist somit eine relative Größe. Daraus ergeben sich rein bildmorphologisch Uneindeutigkeiten in Abhängigkeit von der initialen Verschlusslokalisation bei seiner Verwendung, die sowohl von klinischer Relevanz als auch problematisch beim Training neuronaler Netzwerke sein können. Für einen universellen Einsatz neuronaler Netzwerke wurde die Hypothese aufgestellt, dass sich visuell ermittelte Perfusionsscores, die sich aus dem prozentualen Anteil perfundierten Hirnareals ableiten, in einer mit dem Scoring nach TICI vergleichbaren Interrater-Variabilität bestimmen lassen und sich somit als eine zum TICI alternative *ground truth* beim Training neuronaler Netzwerke eignen (Hypothese 2). Diese Hypothese sollte mit Hilfe des gleichen Datensatzes durch einen stichprobenartigen Vergleich der vergebenen Perfusionsscores zwischen zwei *ratern* untersucht werden.

4 Material und Methoden

4.1 Erhebung des Datensatzes

Die Erhebung des Datensatzes erfolgte retrospektiv mit Hilfe einer klinikinternen, neuroradiologischen Datenbank. Diese Datenbank wird von Ärzten:innen und Mitarbeitern:innen der Klinik für neuroradiologische Diagnostik und Intervention gepflegt und beinhaltet anonymisierte Informationen von allen Schlaganfallpatienten, die seit Oktober 2009 am Universitätsklinikum Hamburg-Eppendorf eine mechanische Thrombektomie erhalten haben. Zum Zeitpunkt der letzten Datenextraktion am 15. August 2018 umfasste die Datenbank 923 Einträge. Das Patientenkollektiv innerhalb der Datenbank ist auf Grund ihrer Eignung zur mechanischen Rekanalisation vorselektioniert, wobei die Verschlüsse in der distalen ACI/Karotis-T, im vorderen sowie im hinteren Kreislauf liegen. Das mittlere Erkrankungsalter der Patienten betrug 69 ± 15 Jahre bei einer ausgeglichenen Geschlechterverteilung. Neben patienten- und therapiebezogenen Informationen fanden sich auch klinische Scores (wie ASPECT und NIHSS) in der Datenbank, wobei nicht alle Informationen für jeden Patienten vorlagen.

Die Thrombektomien erfolgten in der Klinik für neuroradiologische Diagnostik und Intervention des Universitätsklinikum Hamburg-Eppendorf unter leitlinienkonformen Standards und unter Verwendung zweier Bildebenen. Streng standardisierte Parameter zur Akquisition der angiographischen Serien, wie zum Beispiel Bildausschnitt und -winkel, Menge des injizierten Kontrastmittels und Durchleuchtungsdauer der einzelnen Serien existieren bei gegebenem klinischem Vorrang nicht.

4.2 Ein- und Ausschlusskriterien zur Aufnahme in den Datensatz

Die Erhebung des Datensatzes zur Beantwortung von Hypothese 1 ist detailliert in der Publikation beschrieben. Die wichtigsten Inklusionskriterien sind hier kurz wiedergegeben:

- Nur Verschlüsse des M1-Segments der A. cerebri media
- Zwei Projektionsebenen im anterior-posterioren sowie im lateralen Strahlengang
- Stromgebiet zu mindestens $\frac{2}{3}$ dargestellt
- Anteil des Stromgebiets am *field of view* der Serie mindestens 30%
- Mindestens fünf Bilder pro Serie
- Kein körperfremdes Material im betroffenen Stromgebiet einliegend

Zur Beantwortung von Hypothese 2 wurden bei ansonsten analogen Inklusionskriterien zusätzlich zu den für Hypothese 1 verwendeten M1-Verschläüssen auch Karotis-T, M2 und A. cerebri anterior Verschläüsse evaluiert.

Von einer Vorher-nachher-Paarung von Bildserien wurde aus folgenden Gründen abgesehen:

- die Paarung von Serien mit ähnlichen Aufnahmebedingungen schränkt den Datensatz stark ein
- die Definition entsprechender Paare bedarf menschlicher Beteiligung, was mit dem Konzept einer Automatisierung in Widerspruch steht
- das Training neuronaler Netzwerke mit Paaren ist technisch aufwändiger
- die Beschränkung auf M1-Verschläüsse erlaubt eine bildmorphologische Eindeutigkeit bezüglich des TICI Scores aus einer Serie

Für eine nachträgliche Evaluation potentieller Fehlerquellen bei der Prädiktion wurden für alle Angiographieserien bestimmte, zur Darstellung kommende Eigenschaften quantifiziert:

- Venöse Kontrastmittelfase: beginnend erfasst oder nicht erfasst
- Kontrastierung der A. ophtalmica: nicht, schwach oder deutlich erkennbar
- Kontrastierung der A. temporalis anterior: nicht erkennbar oder erkennbar
- Retrograde Perfusion des verschlossenen Stromgebiets über Kollateralen (typischerweise leptomeningeale Kollateralen): nicht erkennbar, leicht oder stark
- Antegrade Perfusion des hinteren Stromkreises über Pcom: nicht erkennbar, leicht oder stark
- Kontrastierung der kontralateralen Hemisphäre über Acom oder Perforansgefäße: nicht erkennbar, leicht oder stark

Abbildung 3 zeigt exemplarisch die Dokumentation einer in den Trainingsdatensatz aufgenommenen Serie im zeitlichen Verlauf, dargestellt an fünf relevanten *frames*. Bei der dargestellten Serie handelt es sich um einen proximalen M1-Verschluß mit mTICI 0 bei ausgeprägter leptomeningealer Kollateralisierung des Mediaterritoriums. Pcom und A. ophtalmica kommen nur schwach zur Darstellung, die kontralaterale Hemisphäre ist nicht kontrastiert. Die Serie ist insgesamt 21 *frames* lang und erfasst die venöse Kontrastmittelfase, zeigt keine Bewegungsartefakte oder einliegendes Fremdmaterial und die *region of interest* macht > 50% des *field of view* aus.

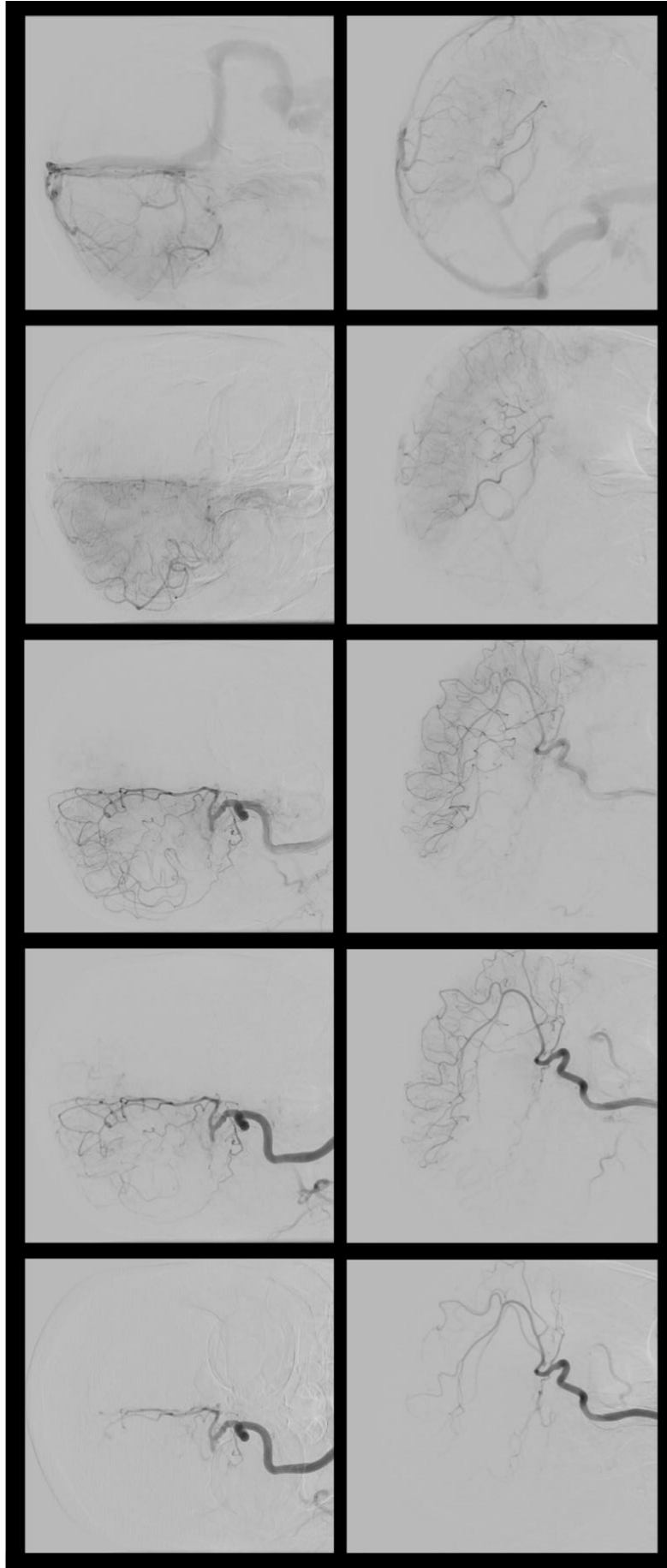


Abbildung 1: Exemplarische Auszüge einer DSA-Serie eines Patienten mit proximalem M1-Verschluss in anterior-posteriorem- und lateralem Strahlengang im zeitlichen Verlauf einer Serie über fünf frames. Das vordere Stromgebiet ist in beiden Ebenen vollständig erfasst, füllt > 50% des FOV und ist gut kontrastiert. Die Serie umfasst insgesamt 21 Bilder, zeigt keine Bewegungsartefakte und es kommen keine Fremdmaterialien intra- wie extrakorporal zur Darstellung. Bei mTICI 0 zeigt sich eine deutliche Kollateralisierung des Medialstromgebiets aus dem Stromgebiet der A. cerebri anterior. Die A. ophthalmica kommt deutlich, die Pcom schwach zur Darstellung, eine Kontrastierung der kontralateralen Hemisphäre erfolgt nicht.

4.3 Externer Datensatz

Zur Validierung der Robustheit des trainierten neuronalen Netzwerks wurde die Prädiktionsgenauigkeit auf einen externen Datensatz wie in der Publikation beschrieben getestet. Die anonymisierten und ungelabelten Daten wurden zunächst unabhängig und später im Konsens von M. Waldmann (*rater*) und einem erfahrenen Neuroradiologen (*senior rater*, > 8 Jahre klinische Erfahrung) nach mTICI gescored. Vor dem Konsensrating erfolgte im zeitlichen Abstand (4 Monate) ein erneutes, verblindetes *rating* der Serien durch den *rater*. Aus diesen *ratings* wurde sowohl eine Intrarater- sowie eine Interrater-Variabilität gemäß Cohens κ bestimmt. Der externe Datensatz wurde in einem ersten Experiment vollständig dem ausschließlich auf den internen Datensatz trainierten *deep learning* Model ohne weitere Anpassung des Models zur Prädiktion vorgelegt. In einem zweiten Experiment wurde das Model mit der einen Hälfte des externen Datensatzes retrainiert, bevor eine Prädiktion auf die andere Hälfte des Datensatzes erfolgte.

4.4 Training und Testung der neuronalen Netzwerke

Zur Beantwortung von Hypothese 1 wurde der Datensatz für Training und Testung der neuronalen Netzwerke an die Mitarbeiter des *Institute for Computational Neuroscience* übergeben und wie in der Publikation beschrieben prozessiert. Training und Testung des neuronalen Netzwerks ist hier kurz beschrieben:

- Der Datensatz wurde in einen Trainings- (82%) und in einen Testdatensatz (18%) aufgeteilt.
- Die Serien wurden auf 224x224 Pixel herunterskaliert und der Intensitätsbereich auf Werte zwischen [-1, 1] normalisiert.
- Die Architektur des *deep learning* Systems bestand aus einer Kombination von speziellen Konvolutionsnetzen, sogenannten GRUs (*gated recurrent units*), mit einer *EfficientNet B0 architecture*.
- Die GRUs dienen der Nutzung der Raum-Zeit aufgelösten Informationen von DSA-Serien, die *EfficientNet B0 architecture* der Fusionierung der Informationen aus den beiden Projektionsebenen.

4.5 Bestimmung der Perfusionsscores

Zur Beantwortung von Hypothese 2 wurden allen inkludierten Bildserien ein zum TICI Score alternatives Label zugeteilt. Visuell erhoben wurden zwei Scores: einer für die antegrade Perfusion des Mediastromgebiets (ACM) sowie einer für die Perfusion des vorderen Kreislaufs (Media- + Anterior-Stromgebiet, *Hemi*) der jeweiligen Hemisphäre. Die Abschätzung des hämodynamischen Beitrags der jeweiligen Stromgebiete an der Gesamtpfusion der entsprechenden Hemisphäre orientierte sich an publizierten Werten, die eine ungefähre Verteilung von $\frac{1}{3}$ zu $\frac{2}{3}$ der Perfusion des vorderen Kreislaufs zwischen Anterior- und Mediastromgebiet beschrieben.⁴⁹

Da eine exakte prozentuale Angabe der Perfusion des entsprechenden Stromgebiets durch visuelle Bestimmung anhand angiographischer Bilddaten nicht möglich ist, wurde eine Kategorisierung in insgesamt neun Perfusionsscores geschaffen (Tabelle 4), in die eine visuelle Einteilung mit guter Reliabilität erwartet wurde. Die Grenzen der Scores wurden so gewählt, dass typische Verschlusskonstellationen sicher einer Klasse zugeordnet werden konnten.

Tabelle 4: Definition der eingeführten Perfusionsscores [ps], welche sich aus angiographischer Bilddaten visuell abgeschätzten Perfusionsprozenten des jeweiligen Stromgebiets ableiten. Mit der Einführung des Perfusionsscores sollte dem Umstand Rechnung getragen werden, dass eine exakte Bestimmung der prozentualen Hirnperfusion eines Stromgebiets visuell nicht möglich ist und so eine höhere Reliabilität durch Reduktion der Bildinformation bei der Quantifizierung erreicht werden.

Perfusionsscore [ps]	Perfundiertes Stromgebiet	Beschreibung
0	0%	Keine Perfusion
1	1 – 5%	Minimale Perfusion eines Teils des Stromgebiets; z. Bsp. bei nicht proximalem M1-Verschluss
2	6 – 20%	Teilweise Perfusion einzelner Äste; z. Bsp. partielle Perfusion einzelner M2-Äste
3	21 – 40%	Teilweise Perfusion mehrerer Äste; z. Bsp. Perfusion eines kleinkalibrigen M2-Truncus
4	41 – 60%	Perfusion ganzer distaler Stromgebiete, z. Bsp. Perfusion eines M2-Truncus
5	61 – 80%	Weitreichende Perfusion mit Verschluss distaler Stromgebiete; z. Bsp. fortbestehender Okklusion eines M3-Gebiets
6	81 – 94%	Weitestgehende Perfusion mit Verschluss einiger distaler Äste
7	95 – 99%	Fast vollständige Perfusion oder verzögertes Kontrastmittelverhalten; z. Bsp. bei Verschluss eines distalen Astes
8	100%	Vollständige Perfusion

Die Beschreibung einer Serie durch zwei Label erlaubt bei vergleichender Betrachtung eine Unterscheidung zwischen distalen ACI/Karotis-T-Verschlässen, bei denen der gesamte vordere Kreislauf nicht perfundiert ist, und reinen Mediaverschlässen, bei denen ein Teil des vorderen Kreislaufs durch die A. cerebri anterior (ACA) perfundiert ist. Die Perfusion des hinteren Kreislaufs blieb auf Grund der unterschiedlich stark ausgebildeten Pcom und der damit einhergehenden unterschiedlich starken Kontrastierung unberücksichtigt. Abbildung 4 zeigt beispielhaft drei typische Verschlusskonstellationen und die dazugehörigen, vergebenen Perfusionsscores für das Mediaterritorium sowie für den vorderen Kreislauf der entsprechenden Hemisphäre, welche auch ohne Kenntnis der Bilddaten Rückschlüsse auf das Verschlussmuster erlauben.

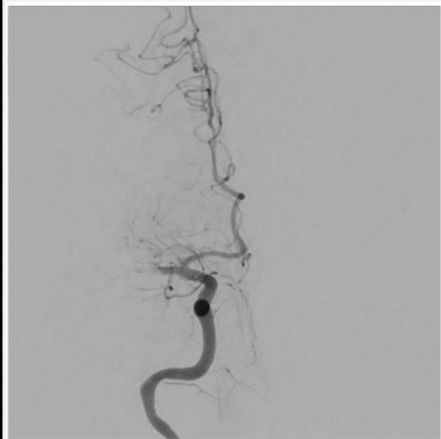
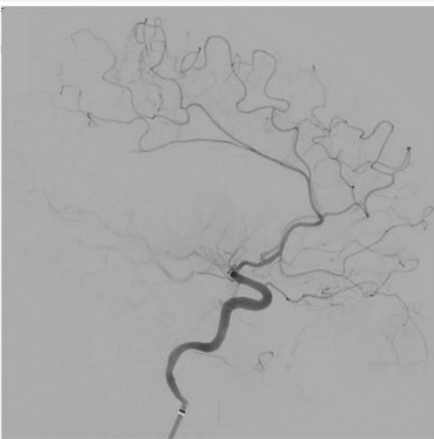
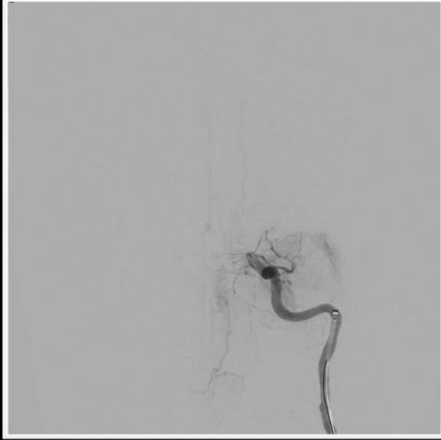
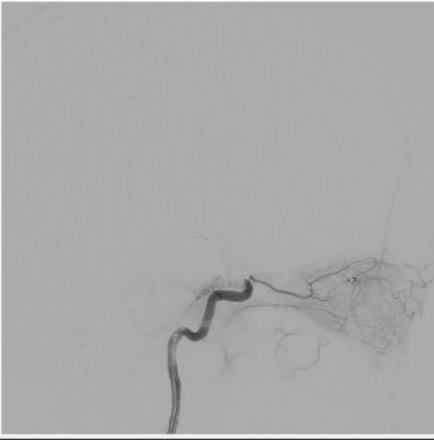
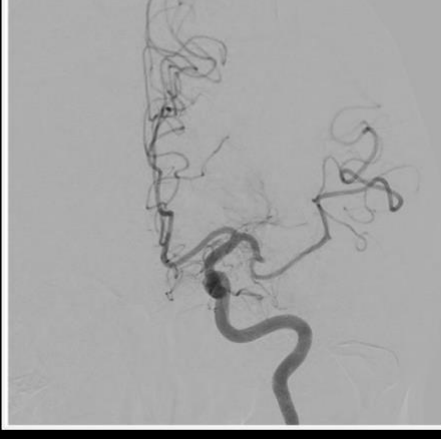
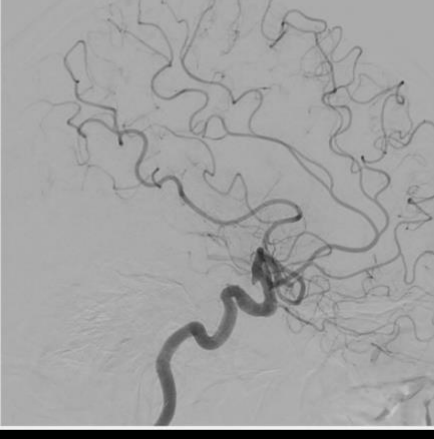
anterior-posterior	lateral	Perfusionsscore	
		MCA 1	Hemi 3
		MCA 0	Hemi 0
		MCA 3	Hemi 5

Abbildung 4: Identifikation des Verschlussmusters mit Hilfe der vergebenen Perfusionsscores. Ist der score für die Hemisphäre größer als der für das Mediastromgebiet, muss das Anteriorstromgebiet zumindest teilweise perfundiert sein und ein Mediaverschluss vorliegen (oben). Bei einem score von ps 0 für die gesamte Hemisphäre muss es sich um einen T- bzw. ACI-Verschluss handeln (Mitte). Eine kräftige A. temporalis anterior versorgt einen Teil des Mediastromgebiets (unten).

4.7 Statistische Auswertung der Interrater-Reliabilität für die Perfusionsscores

Die Interrater-Reliabilität bei der Bestimmung des Perfusionsscores wurde stichprobenartig (40 Serien, entspricht 6%) durch Vergleich des Scorings mit einem erfahrenen Neuroradiologen bestimmt. Bei der Auswahl der Serien wurde auf eine ungefähre Gleichverteilung der Häufigkeiten der unterschiedlichen Perfusionsscores geachtet. Die statistische Auswertung nach Cohens κ erfolgte mit Hilfe der Software IBM SPSS (Version 27.0.0). Für die Berechnung des gewichteten Cohens κ wurde die SPSS Erweiterung STATS_WEIGHTED_KAPPA 1.2.1 genutzt und eine lineare Gewichtung verwendet.

5 Ergebnisse

5.1 Hypothese 1: Eignung neuronaler Netzwerke zur automatischen Quantifizierung des Behandlungserfolgs der MT nach TICI

5.1.1 Erhebung des *in-house* Datensatzes

Die Erhebung des *in-house* Datensatzes ist in der Publikation beschrieben. Wesentliche Eckpunkte sind im Folgenden kurz beschrieben:

- 789 Patienten wurden evaluiert.
- Davon zeigten 371 Patienten einen M1-Verschluss, wovon wiederum 236 Patienten die weiteren Inklusionskriterien erfüllten.
- Diese 236 Patienten waren im Schnitt 74 Jahre alt, 55% waren weiblich.
- Von diesen 236 Patienten konnten 491 DSA-Serien in den Datensatz aufgenommen werden, welcher in einen Trainingsdatensatz (401 Serien) und in einen Testdatensatz (90 Serien) aufgeteilt wurde.
- Die Häufigkeiten der verschiedenen TICI Scores sind

mTICI Score	rel. Häufigkeit	abs. Häufigkeit
T0	35%	173/491 Serien
T1	5%	27/491 Serien
T2a	15%	74/491 Serien
T2b	23%	111/491 Serien
T3	22%	106/491 Serien

5.1.2 Externer Datensatz: Bestimmung von Intra- und Interrater-Reliabilität

Zum qualitativen Vergleich der Prädiktionsgenauigkeit des neuronalen Netzwerks auf den externen Datensatz wurde die Inter- und Intrarater-Reliabilität zwischen zwei menschlichen *ratern* auf diesen Datensatz bestimmt. Der externe Datensatz umfasste insgesamt 95 Serien, welche wie in 4.3 beschrieben insgesamt viermal (*rater_{t1}*, *rater_{t2}*, *senior rater* und Konsensrating) geratet wurden. Diese *ratings* wurden in den folgenden sechs Konstellationen statistisch untersucht:

- *Rater_{t1}* gegen *rater_{t2}*
- *Rater_{t1}* gegen *senior rater*
- *Rater_{t2}* gegen *senior rater*

- $Rater_{t1}$ gegen Konsensrating
- $Rater_{t2}$ gegen Konsensrating
- *Senior rater* gegen Konsensrating

Die Ergebnisse der statistischen Analyse zur Intra- und Interrater-Reliabilität zwischen den beiden *ratern* sind in Tabelle 5, der Vergleich mit dem Konsensrating in Tabelle 6 zusammengefasst. Eine exzellente Übereinstimmung findet sich zwischen *senior rater* und dem Konsensrating mit einem Cohens κ von 0.92. Die Übereinstimmungen zwischen den *ratern*, bzw. zwischen *rater* zum Zeitpunkt t_1 und Zeitpunkt t_2 ohne Berücksichtigung des Konsensrating ist mit Werten für Cohens κ zwischen 0.74 und 0.86 ebenfalls gut bis sehr gut und besser als die in der Literatur beschriebenen Werte (siehe Tabelle 3) zur Interrater-Reliabilität.

Tabelle 5: Intra- und Interrater-Reliabilitäten zwischen den beiden ratern untereinander ohne Berücksichtigung des Konsensrating.

	$rater_{t1}/rater_{t2}$	$rater_{t1}/senior\ rater$	$rater_{t2}/senior\ rater$
Übereinstimmung	89%	82%	80%
Cohens κ	0.86	0.77	0.74

Tabelle 6: Interrater-Reliabilitäten zwischen den beiden ratern im Vergleich mit dem Konsensrating.

	$rater_{t1}/Konsens$	$rater_{t2}/Konsens$	<i>senior rater</i> /Konsens
Übereinstimmung	87%	85%	94%
Cohens κ	0.83	0.81	0.92

5.1.3 Statistische Analyse der Prädiktionsgenauigkeit der NN auf die verschiedenen Datensätze

Die Ergebnisse zur Klärung der Frage nach der Eignung von neuronalen Netzwerken zur automatisierten Quantifizierung des Behandlungserfolgs nach MT sind detailliert in der Publikation dargestellt. Die wichtigsten Ergebnisse sind in Tabelle 7 kurz wiedergegeben. Für die Auswertung wurden die mTICI Klassen mTICI 0 und mTICI 1 zu einer Kategorie zusammengefasst, da mTICI 1 zum einen auf Grund seiner relativen Seltenheit im Model deutlich unterrepräsentiert ist und zum anderen auch in der Literatur gängig als verschlossener Zustand betrachtet wird. Es zeigt sich, dass auf den *in-house* Testsatz eine Übereinstimmung von 71% mit einem Cohens κ von 0.61 erreicht wurde, was einer guten Übereinstimmung entspricht. Auf den externen Datensatz fiel die Übereinstimmung mit 64%

schlechter aus, konnte aber durch eine Anpassung des Models mit der Hälfte der Serien des externen Datensatzes auf die verbliebene Hälfte auf eine Übereinstimmung von 74% gesteigert werden.

Tabelle 7: Ergebnisse der Prädiktion des neuronalen Netzwerks auf den in-house Testsatz, auf den vollständigen externen Datensatz ohne vorherige Anpassung des DL-Models, sowie auf den mit der Hälfte des externen Datensatz retrainierten DL-Models. Angegeben ist die Übereinstimmung, die sliding-window Übereinstimmung_{sw} sowie Cohens κ .

	In-house Testsatz	Externer Datensatz (Model nicht angepasst)	Externer Datensatz (Model angepasst)
Übereinstimmung	71%	64%	74%
Übereinstimmung_{sw}	100%	97%	97%
Cohens κ	0.61	0.52	0.66

In dem *in-house* Testsatz zeigten elf Serien eine nahezu vollständige Reperfusion von 90-99%, wovon sechs Serien durch das Netzwerk korrekt als mTICI 2b prädiziert wurden. Zehn Serien des Testsatzes stellten einen Grenzfall zwischen mTICI 2a und 2b mit einer Reperfusion zwischen 40 – 60% dar, wovon wiederum sieben Serien durch das Netzwerk übereinstimmend mit der angelegten *ground truth* bewertet wurden. Dies zeigt zum einen, dass TICI Scores zwischen zwei Klassen - ähnlich wie es zwischen zwei menschlichen *ratern* zu erwarten ist - häufiger falsch zugeordnet werden, aber auch, dass sie durch das Netzwerk zu knapp $\frac{2}{3}$ übereinstimmend zu der *ground truth* klassifiziert werden. Weitere charakteristische, bildmorphologische Häufungen, die eine systematische Erklärung für eine Abweichung hätten geben können, wie zum Beispiel eine ausgeprägte Perfusion der kontralateralen Hemisphäre oder des hinteren Kreislaufs, mäßige Bewegungsartefakte oder ein geringer Kontrast, konnten nicht identifiziert werden. Für den externen Datensatz ergab sich als generelle Auffälligkeit, dass die Bildakquisition erst zu einem recht späten Zeitpunkt nach Kontrastmittelapplikation begonnen wurde, so dass das Anfluten des Kontrastmittels im Vergleich zu dem *in-house* Datensatz häufig nicht abgebildet wurde. Für beide Datensätze ist die Vorhersagegenauigkeit für die mTICI Scores T0/1 und T3 am höchsten, für mTICI T2b am niedrigsten.

Abbildung 5 zeigt Beispiele für korrekte Prädiktionen des DL-Models auf den *in-house* Testsatz exemplarisch an schwierigen Perfusionskonstellationen wie bei einer ausgeprägten Kollateralisierung oder wie bei Grenzfällen zwischen zwei mTICI Scores. Abbildung 6 zeigt exemplarisch falsche Prädiktionen auf den *in-house* Testsatz anhand zweier Grenzfälle.

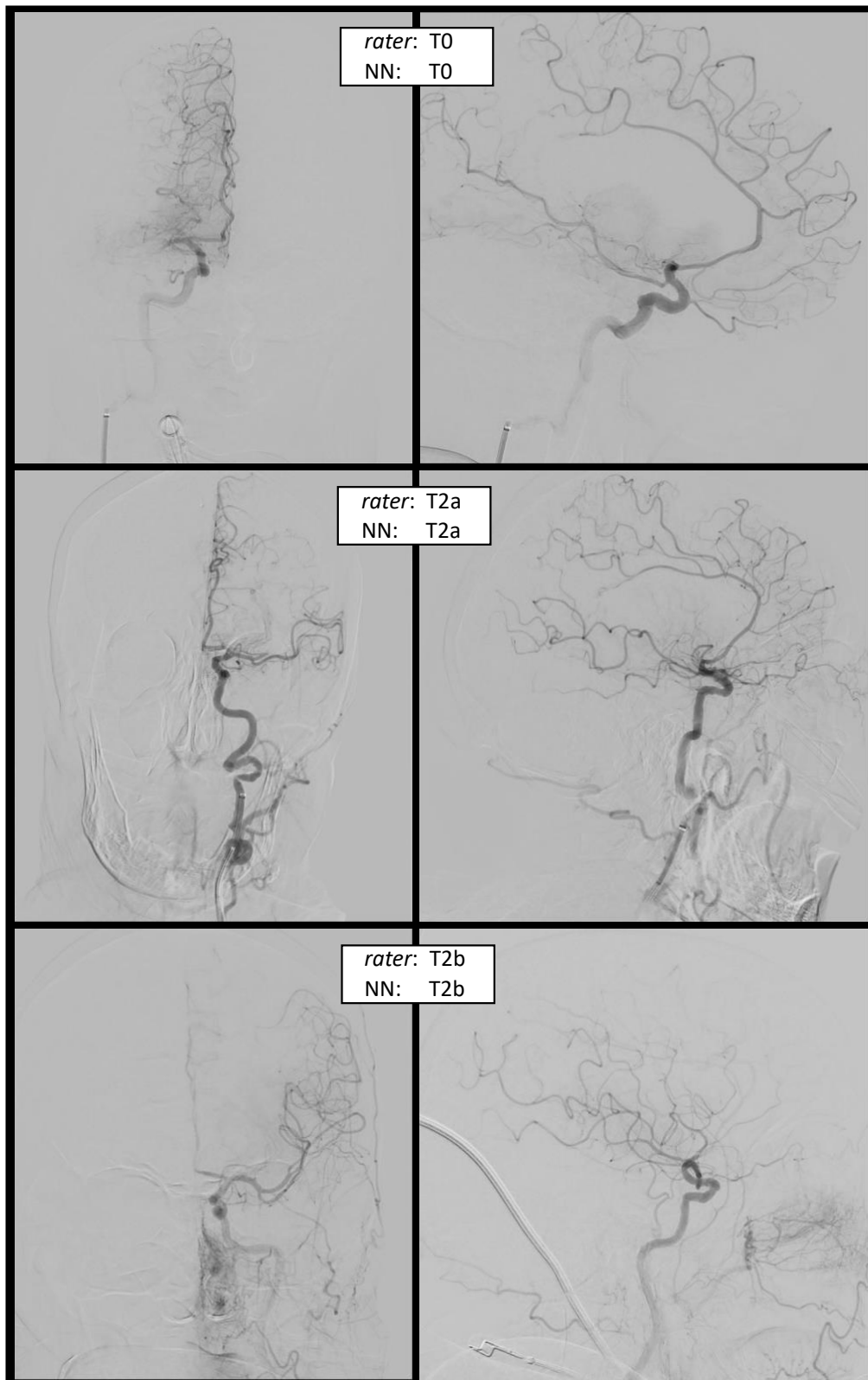


Abbildung 5: Beispiele für korrekte Prädiktionen des neuronalen Netzwerks (NN) auf Serien aus dem in-house Testdatensatz. Gezeigt ist die anterior-posteriore (links) und die zu dem frame gehörende laterale Projektionsebene (rechts). Oben: Korrekte Zuordnung in die Klasse mTICI 0 bei ausgeprägte leptomeningealer Kollateralisierung des Medialstromgebiets und einer ausgeprägten Pcom. Mitte: Korrekte Zuordnung in die mTICI Klasse 2a bei perfundiertem Truncus inferior. Unten: Korrekte Zuordnung in die Klasse 2b bei beinahe perfekter Rekanalisation. Einzelne Astabbrüche und eine verzögerte Perfusion in der vorderen Astgruppe werden in der zeitaufgelösten Betrachtung erkennbar.

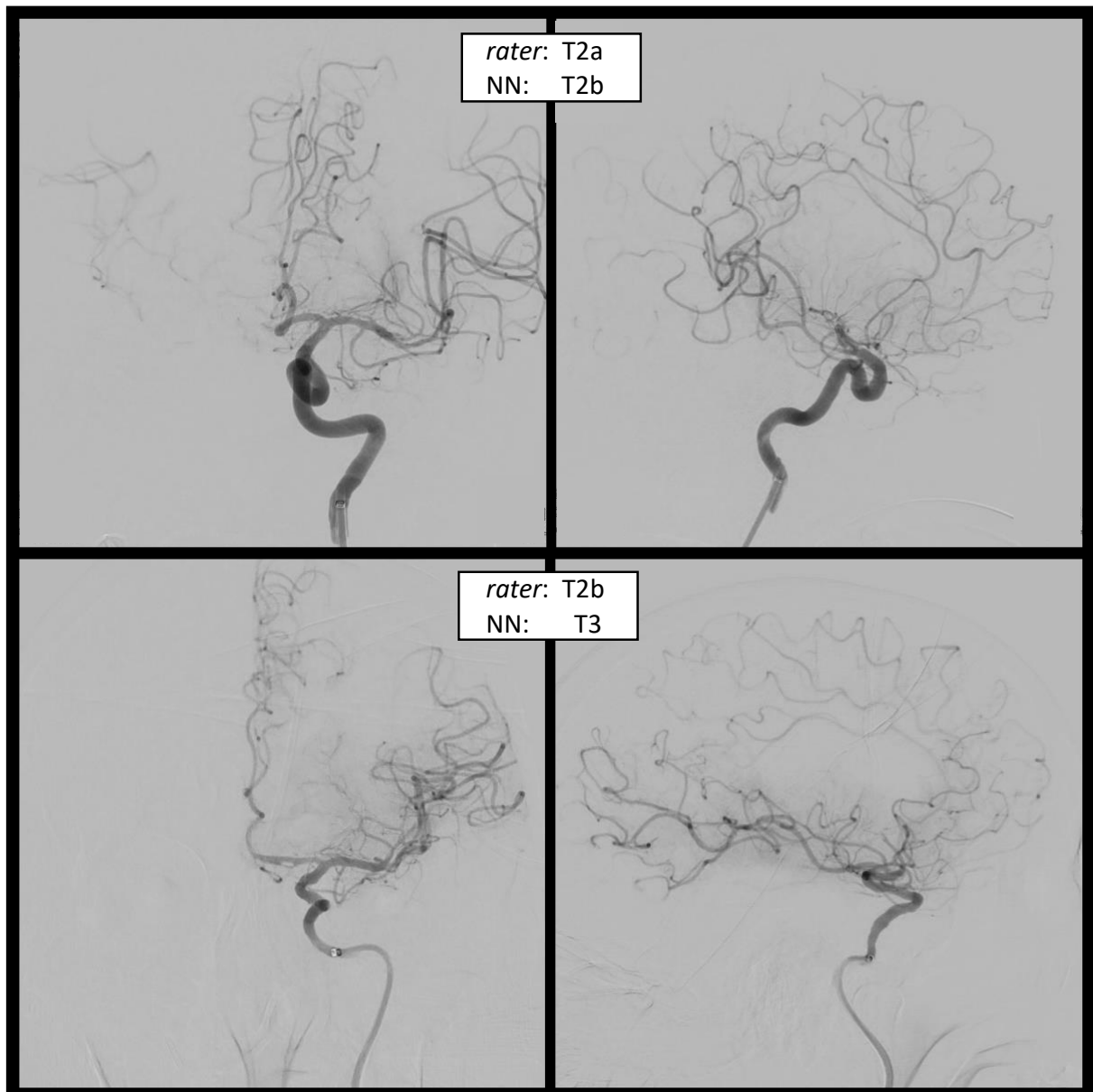


Abbildung 6: Beispiele für falsche Prädiktionen des neuronalen Netzwerks (NN) auf Serien aus dem in-house Testdatensatz. Gezeigt ist die anteriore-posteriore (links) und die zu dem frame gehörende laterale Projektionsebene (rechts). Oben: Im Konsens wurde für die Serie ein mTICI 2a als ground truth bestimmt, wobei dies bei perfundiertem, kräftigen Truncus inferior und verschlossener zentraler und frontaler Astgruppe eine diskutable Entscheidung ist. Die Netzwerk Prädiktion für die Serie lautet mTICI 2b. Unten: Bei teilverschlossenem Truncus superior mit entsprechendem Perfusionsdefizit bzw. Perfusionsverzögerung vor allem in der frontalen Astgruppe wurde als ground truth mTICI 2b bestimmt. Die Prädiktion des Netzwerks war mTICI 3.

5.1.4 Einordnung der Prädiktionsgenauigkeit mit in der Literatur beschriebenen Werten zur Interrater-Reliabilität

Die statistische Auswertung der Prädiktionsgenauigkeit auf den *in-house* Testsatz ergab mit einem Cohens κ von 0.61 nicht nur einen Wert, der gängig als substantielle bis gute Übereinstimmung interpretiert wird, sondern auch einen Wert, der im Vergleich mit den in

der Literatur beschriebenen Werten (siehe Tabelle 3) eine Prädiktion auf menschlichem Niveau zeigt. Sowohl die Übereinstimmung von 71% auf alle Serien des *in-house* Testsatzes als auch die in Kapitel 5.1.3 genannte Übereinstimmung von 61% auf Serien mit einer Perfusion im Grenzbereich zwischen zwei TICI Scores liegt über der in der groß angelegten Studie von Zhang et al. ermittelten Übereinstimmung von 56% zwischen Interventionalist vor Ort und *core facility*.³⁴ Die Ergebnisse auf den externen Datensatz waren nach Anpassung des Models mit nur einer kleinen Serienzahl sogar noch besser als die Ergebnisse auf den *in-house* Datensatz, was dafür spricht, dass neuronale Netzwerke zur Beurteilung von Thrombektomieserien nach TICI geeignet sind.

5.2 Hypothese 2: Perfusionsscores als Alternative zum TICI Score

5.2.1 Bestimmung der Perfusionsscores

Zur Klärung der Frage, ob sich visuell ermittelte Perfusionsscores als zum TICI Score alternatives Label eignen, wurden insgesamt 678 Serien von 258 Patienten ein Perfusionsscore zugeordnet. Für eine eventuell spätere Verwendung des Scores zum Training neuronaler Netzwerke wurden zur Beantwortung von Hypothese 1 analoge Inklusionskriterien angelegt und der Datensatz um höhergradige Mediaverschlüsse sowie um Karotis-T-Verschlüsse ergänzt.

Tabelle 8 zeigt die Häufigkeiten der vergebenen Scores für das Mediastromgebiet (horizontal verlaufend) sowie für die Hemisphäre (vertikal verlaufend) in Form einer Kreuztabelle. In der rechten, grau hinterlegten Spalte ist die Summe der Häufigkeit des jeweiligen Perfusionsscores für das Mediastromgebiet und in der untersten, hellgrau hinterlegten Zeile die für den gesamten vorderen Kreislauf wiedergegeben. In der darüber liegenden Matrix sind die Häufigkeiten der verschiedenen Kombinationen der Perfusionsscores aus dem Media- und gesamten vorderen Kreislauf dargestellt. Aus den Kombinationen lassen sich Verschlusslokalisationen ableiten, wobei sich vereinfachend sagen lässt, dass unterhalb der gedachten Diagonale zwischen „keine Perfusion“ und „vollständige Perfusion“ eher ein Verschluss von Ästen der A. cerebri anterior und oberhalb der Diagonale eher ein Verschluss von Ästen der A. cerebri media führend ist. Farblich hervorgehoben sind einige typische Verschlusslokalisationen wie T (gelb), M1 (grün), A. cerebri media + A. cerebri anterior kombiniert (rosa) oder ein persistierendes Perfusionsdefizit im Stromgebiet der ACA bei vollständiger Wiedereröffnung des Mediastromgebiets (orange). Kombinationen von sich in

sich ausschließenden Scores, wie zum Beispiel ein hoher Score für den vorderen Kreislauf in Kombination mit einem niedrigen Score für das Mediastromgebiet traten nicht auf.

Tabelle 8: Kreuztabelle mit den Häufigkeiten der vergebenen Perfusionsscores für das Mediastromgebiet (horizontal verlaufend) sowie für den gesamten vorderen Kreislauf (vertikal verlaufend). Unterste Zeile, hellgrau: Summe der Häufigkeit der verschiedenen Perfusionsscores bezogen auf den vorderen Kreislauf. Rechte Spalte, dunkelgrau: Häufigkeit der verschiedenen Perfusionsscores bezogen auf das Mediastromgebiet. Typische Verschlusskonstellationen sind farbig hervorgehoben. Gelb: T-Verschluss. Grün: M1-Verschluss, wobei hellgrün einem proximalen und dunkelgrün eher einem distalen M1-Verschluss entspricht. Orange: Vollständige M1-Reperfusion mit Perfusionsdefizit im Stromgebiet der A. cerebri anterior. Rosa: Perfusionsdefizit führend in der A. cerebri anterior. Hellblau: vollständige Rekanalisation des vorderen Stromgebiets.

		Vorderer Kreislauf									Σ ACM
		ps 0	ps 1	ps 2	ps 3	ps 4	ps 5	ps 6	ps 7	ps 8	
Mediastromgebiet	ps 0	50	2	3	31	-	-	-	-	-	86
	ps 1	-	3	6	165	-	-	-	-	-	174
	ps 2	-	-	6	29	27	-	-	-	-	62
	ps 3	-	-	1	-	21	7	-	-	-	29
	ps 4	-	-	-	1	5	51	5	-	-	62
	ps 5	-	-	-	-	-	9	21	-	-	30
	ps 6	-	-	-	-	2	2	20	27	-	51
	ps 7	-	-	-	-	-	4	2	51	-	57
	ps 8	-	-	-	-	-	5	5	1	116	127
Σ vorderer Kreislauf		50	5	16	226	55	78	53	79	116	678

Die Vergabe der Perfusionsscores war trotz der feinen Graduierung in insgesamt neun Kategorien mit hoher Kohärenz möglich, da hier, anders als beim TICI, bestimmte Gefäß- und Verschlusskonstellationen wie zum Beispiel eine ausgeprägte A. temporalis anterior, eine geringe Restperfusion oder Thrombuslokalisation im Übergang von Gefäßabschnitten valider berücksichtigt werden konnten. Angiographische Beispiele für die verschiedenen Perfusionsscores bezogen auf das Mediastromgebiet sind in Abbildung 7 und bezogen auf den gesamten vorderen Kreislauf in Abbildung 8 dargestellt.

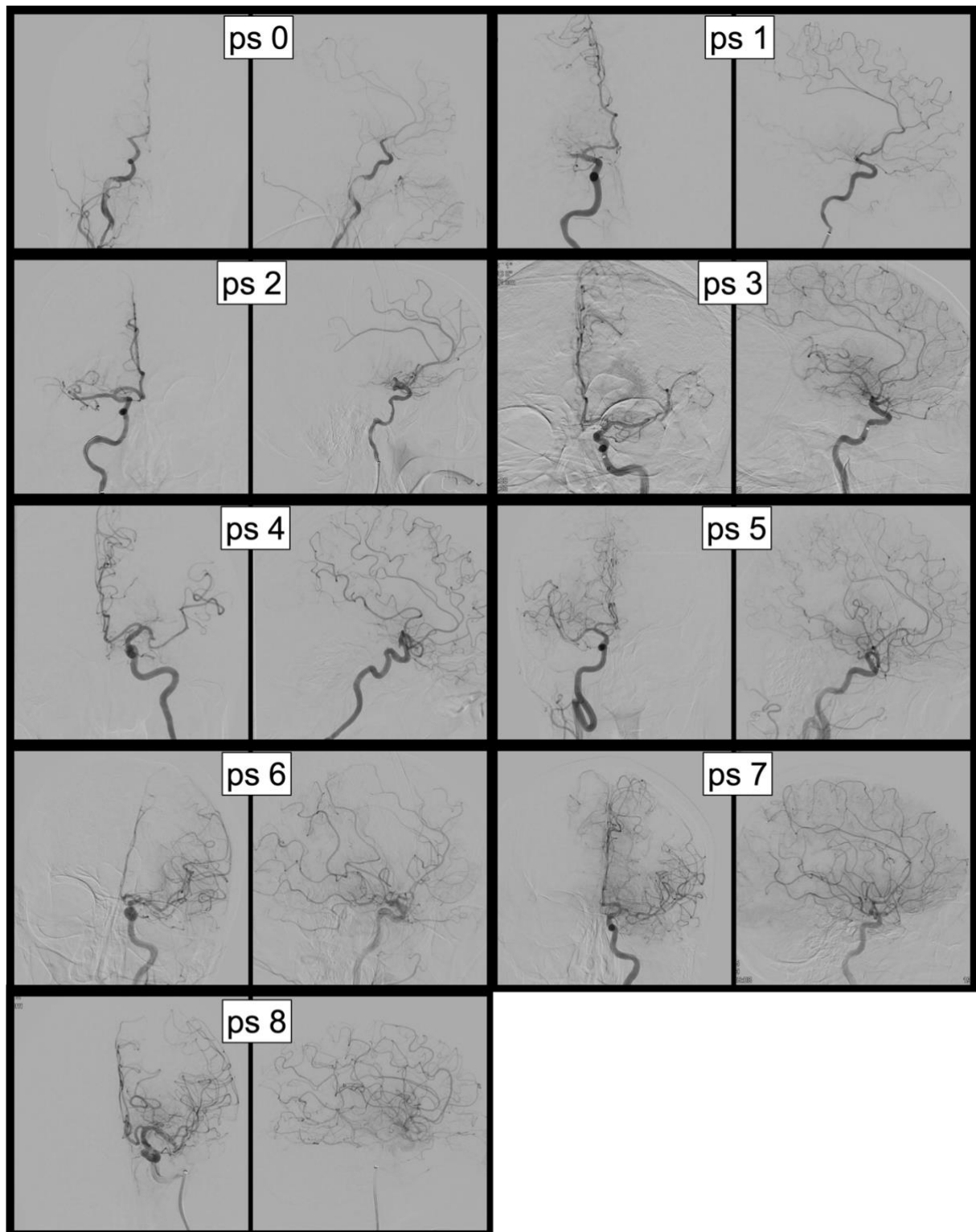


Abbildung 7: Exemplarische Darstellung von DSA-Serien für die verschiedenen Perfusionsscores in anterior-posterioren sowie in lateralen Strahlengang bezogen auf das Stromgebiet der A. cerebri media. Gezeigt sind jeweils die Bildpaare mit der maximalen arteriellen Kontrastierung der cerebralen Gefäße. Zwischen den zusammengehörenden Projektionsebenen ist der visuell bestimmte Perfusionsscore ps für die jeweilige Serie (ps 0 = 0% Perfusion, ps 1 = 1 - 5% Perfusion, ps 2 = 6 - 20% Perfusion, ps 3 = 21 - 40% Perfusion, ps 4 = 41 - 60% Perfusion, ps 5 = 61 - 80% Perfusion, ps 6 = 81 - 94% Perfusion, ps 7 = 95 - 99% Perfusion, ps 8 = 100% Perfusion).

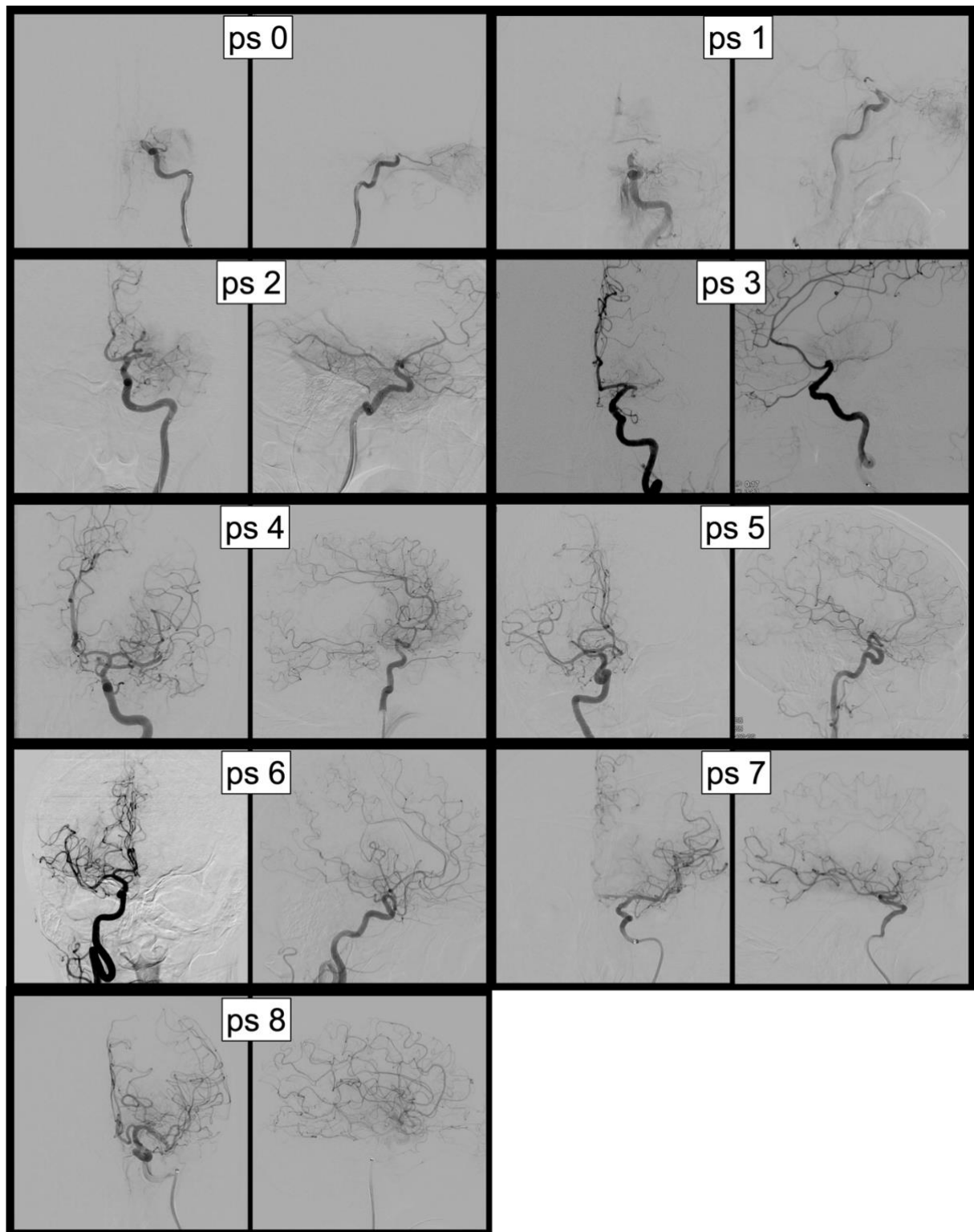


Abbildung 8: (Abbildung Mo 3) Exemplarische Darstellung von DSA-Serien für die verschiedenen Perfusionsscores in anterior-posteroren und lateralen Strahlengang bezogen auf den gesamten vorderen Kreislauf. Gezeigt sind jeweils die Bildpaare mit der maximalen arteriellen Kontrastierung der cerebralen Gefäße. Zwischen den zusammengehörenden Projektionsebenen ist der visuell bestimmte Perfusionsscore ps für die jeweilige Serie (ps 0 = 0% Perfusion, ps 1 = 1 - 5% Perfusion, ps 2 = 6 - 20% Perfusion, ps 3 = 21 - 40% Perfusion, ps 4 = 41 - 60% Perfusion, ps 5 = 61 - 80% Perfusion, ps 6 = 81 - 94% Perfusion, ps 7 = 95 - 99% Perfusion, ps 8 = 100% Perfusion).

5.2.2 Bestimmung der Interrater-Reliabilität

An insgesamt 40 der 678 inkludierten DSA-Serien wurde die Interrater-Reliabilität bei der Zuordnung der Perfusionsscores durch Vergleich mit einem erfahrenen Neuroradiologen bestimmt (*senior rater*). Durch diesen wurden ebenfalls Perfusionsscores sowohl für das Mediastromgebiet sowie für den vorderen Kreislauf bestimmt. Die Ergebnisse für die Abweichung beim Perfusionsscore zwischen *rater* (ps_R) und *senior rater* (ps_{SR}) sind in Tabelle 9 dargestellt.

Tabelle 9: Bestimmung der Interrater-Reliabilität bei der Ermittlung des Perfusionsscores stichprobenhaft an insgesamt 40 der 678 inkludierten DSA-Serien sowohl für das Mediastromgebiet als auch für den gesamten vorderen Kreislauf. Für beide Fälle konnte an 24 Serien (60%) eine Übereinstimmung erreicht werden (grün hinterlegt), bei 14 Serien erfolgte die Zuordnung in eine benachbarte Klasse (gelb hinterlegt), was einer sliding-window Übereinstimmung von 95% entspricht. In nur zwei Fällen erfolgte die Zuordnung in die übernächste Klasse (orange hinterlegt).

$\Delta = ps_{SR} - ps_R$	ACM	Vorderer Kreislauf
0	24	24
1	3	2
-1	11	12
2	0	0
-2	2	2
Σ	40	40
$\emptyset \Delta$	0.45	0.45
Übereinstimmung	60%	60%
Übereinstimmung_{sw}	95%	95%
Cohens κ	0.54 (SD 0.08)	0.53 (SD 0.09)
Cohens $\kappa_{\text{gewichtet}}$	0.84 (SD 0.04)	0.80 (SD 0.05)

Die Auswertung der Interrater-Reliabilität erfolgte nach Cohens κ erfolgte sowohl ungewichtet als auch mit einer linearen Wichtung der ordinalskalierten Perfusionsscores. Bezogen auf beide gescorten Kategorien stimmte in 24 von 40 Serien der Perfusionsscore zwischen *rater* und *senior rater* überein und in 14 Fällen betrug die Abweichung ± 1 . Damit konnte in 60% der Fälle eine Übereinstimmung und in 95% der Fälle eine *sliding-window* Übereinstimmung (Übereinstimmung + benachbarte Klassen) erreicht werden. In 5% der Fälle betrug die Abweichung mehr als eine Klasse, wobei die Abweichung nie mehr als zwei Klassen entsprach. Erwartungsgemäß war die Übereinstimmung in den Gruppen der Perfusionsscores ps_0 , ps_1 , ps_2 , ps_3 und ps_8 am höchsten.

6 Diskussion

6.1 Hypothese 1: Eignung neuronaler Netzwerke zur automatischen Quantifizierung des Behandlungserfolgs der MT nach TICl

Eine Diskussion der Ergebnisse findet sich in der Originalpublikation. Eine kurze Einordnung der Ergebnisse folgt hier.

Die vorliegende Arbeit ist nicht nur die erste Studie, die sich mit der Klärung der Frage nach der Eignung neuronaler Netzwerke zur automatischen Befundung von DSA-Serien nach MT beschäftigt, sondern auch die erste, die zeitaufgelöste Angiographieserien, welche während der Schlaganfallbehandlung akquiriert wurden, als Datengrundlage für neuronale Netzwerke nutzt. Insgesamt wurden dem DL-System in drei Experimenten 242 unbekannte DSA-Serien zur Prädiktion vorgelegt. Verglichen mit den in Kapitel 2.3 zitierten Studien handelt es sich dabei bezüglich der bewerteten Serienzahl um eine große und somit repräsentative Studie. Lediglich in der Studie von Zhang. *et al.* wurden mehr Serien (1182) durch zwei unabhängige Institutionen (Interventionalist gegen *core facility*) bewertet, wobei eine Übereinstimmung bei dem TICl Score in nur in 56% der Fälle beobachtet wurde.³⁴ Speziell im Vergleich mit dieser Studie zeigen die erhaltenen Übereinstimmungen zwischen 64 – 74% aus den drei Experimenten, dass das entwickelte DL-Model sehr gut in der Lage ist, DSA-Serien von mechanischen Thrombektomien auf menschlichem Niveau nach TICl zu bewerten. Auch im Vergleich mit den anderen in Kapitel 2.3 beschriebenen Studien zur Interrater-Reliabilität sprechen die bestimmten Werte für Cohens κ für ein Scoring auf menschlichem Niveau und damit für die Eignung der neuronalen Netzwerke. Der initiale Abfall der Prädiktionsgenauigkeit auf den externen Datensatz lässt sich durch den systematischen Unterschied des späten Bildakquisitionsbeginns der Serien des externen Datensatz zurückführen, wobei im Vergleich zu den Serien des *in-house* Datensatzes das Anfluten des Kontrastmittels häufig nicht zur Darstellung kommt. Die zunächst schlechtere Performance auf einen externen Datensatz zeigt ein gängiges Problem von *deep learning* Ansätzen: der Überanpassung (overfitting) auf den Trainingsdatensatz und konnte durch ein retrainieren des Models mit nur einer kleinen Anzahl an Trainingsserien mit etwas unterschiedlichen Akquisitionsbedingungen einfach behoben werden. Das bessere Ergebnis für Cohens κ auf den externen Datensatz nach Anpassung des DL-Models auf diesen zeigt auch, dass eine kontinuierliche Optimierung des DL-Systems möglich ist und spricht für eine

linikübergreifende Anwendbarkeit, wenn man das Model mit ein paar weiteren externen Datensätzen, welche leicht voneinander abweichende Akquisitionsbedingungen zeigen, trainiert.

Eine wesentliche Limitation bei dem Training von neuronalen Netzwerken mit nach TICI gelabelten DSA-Serien stellt die Belastbarkeit der angelegten *ground truth* dar, da der visuellen Beurteilung eine subjektive Einschätzung zu Grunde liegt. Dies zeigt sich auch bei der in der Literatur beschriebenen hohen Interrater-Variabilität beim Scoring nach TICI, welche trotz der groben Graduierung des Scores relativ hoch ist. Hypothetische Ursachen, die auf Grund einer individuell unterschiedlichen Beurteilung zu einer geringen Übereinstimmung auch zwischen zwei Menschen führen können, könnten sein:

- Anatomische Varianten, wie zum Beispiel ein akzessorischer M1-Ast, der Teile des Mediastromgebiets (partiell) perfundiert.
- Häufige Verschlusskonstellationen wie zum Beispiel ein verschlossener M2-Truncus, bei denen der TICI Score im Grenzbereich zwischen zwei Klassen liegt.
- Eine klinisch getriebene und damit nicht streng der Definition folgende Auslegung des TICI Scores, bei der zum Beispiel die Einordnung in mTICI 2a oder mTICI 2b nach der klinisch relevanten Frage nach der Fortführung bzw. Beendigung der Therapie erfolgt.
- Unterschiedliche Auslegung der unscharf definierten TICI-Klassengrenzen zwischen mTICI 0, mTICI 1 und mTICI 2a.

Gerade bei den Grenzfällen zwischen mTICI 2a und 2b, von denen das Netzwerk vier von zehn abweichend von der *ground truth* zugeordnet hat, ist davon auszugehen, dass selbst unter erfahrenen Neuroradiologen die Inter- als auch die Intrarater-Variabilität vergleichbar hoch ist.

Nichtsdestotrotz erlaubt das trainierte Netzwerk eine höhere Vergleichbarkeit der vergebenen Scores als es durch *core facilities*, einem Konsensrating oder durch ein Rerating erzielt werden kann, da die Bewertungskriterien des Algorithmus sowohl über die Zeit als auch über den Raum hinweg konstant sind. In einem klinischen Setting wäre eine Implementation einer solchen Funktion bei aktuell einer Berechnungszeit von 3.7 s auf einem Standard CPU in die Angiographieanlage als Entscheidungshilfe und Absicherung für den Interventionalisten vor allem unter suboptimalen Bedingungen denkbar.

6.2 Hypothese 2: Perfusionsscores als zum TICI Score potentiell alternatives Label

6.2.1 Bestimmung der Perfusionsscores

Je zwei Scores (Mediastromgebiet und vorderer Kreislauf) wurden für 678 Serien erhoben. Elegant an der Angabe eines absoluten Scores ist, dass anatomische Varianten und Besonderheiten im Scoring mit höherer Sicherheit berücksichtigt werden können, da eine genaue Zuordnung von Gefäßabschnitten zu definierten Segmenten nicht nötig ist. So ist zum Beispiel eine getrennte Trifurkation des M1-Segments bei einer Okklusion distal des ersten Abgangs nur schwierig mit hoher Reliabilität nach TICI zu scoren. Beispiele für anatomische Varianten, die beim *rating* nach TICI abweichende Scores zwischen *rater* und *senior rater* ergaben, sind in Abbildung 9 gezeigt. Für die gezeigten Fälle ist eine Zuordnung zu einem Perfusionsscore mit höherer Reliabilität wahrscheinlich, was die potentielle Eignung des Perfusionsscores als *ground truth* zum Training neuronaler Netzwerke impliziert.

Weiterer Vorteil der neu eingeführten Perfusionsscores ist, dass es sich dabei im Gegensatz zum TICI Score um eine absolute und nicht um eine zur initialen Verschlusslokalisation relative Größe handelt. Als solche könnten sich die Perfusionsscores bei einem Training neuronaler Netzwerke als *ground truth* sogar besser eignen als der relative TICI Score.

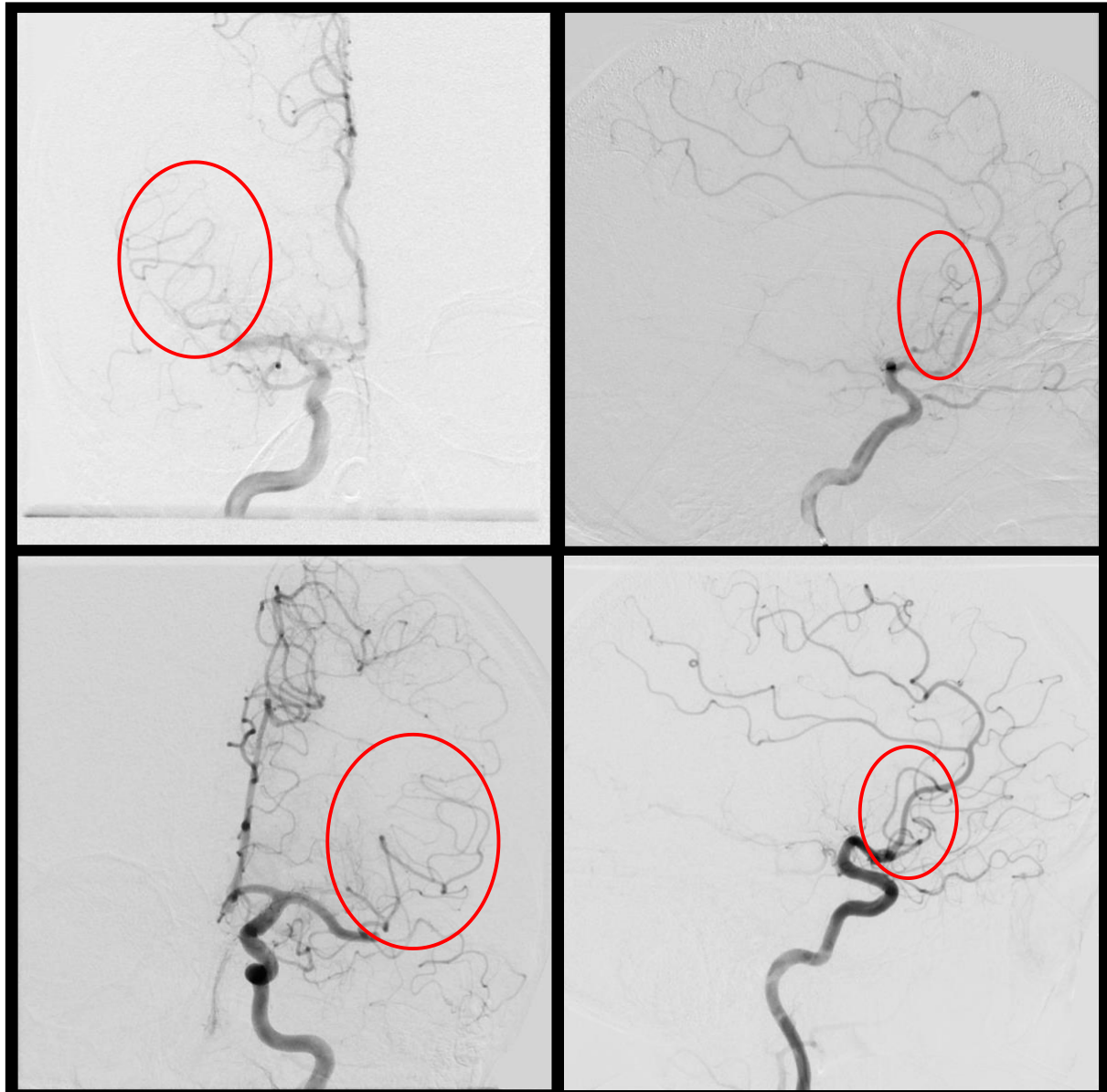


Abbildung 9: Zwei Beispiele (oben und unten) für anatomische Varianten, die zwischen dem *rater* und *senior rater* abweichende Scores ergaben. Gezeigt ist die anterior-posteriore (links) sowie die zu dem *frame* gehörende laterale Projektionsebene. In beiden Fällen erkennt man, dass ein atypischer M1-Ast Teile des vorderen Stromgebiets perfundiert (rote Kreise). Ohne Kenntnis der genauen initialen Thrombuslokalisation bei M1-Verschluss variierte hier die Beurteilung durch die *rater* zwischen mTICI 0 und mTICI 2a, da die Interpretation des Astes unterschiedlich war. Eine Zuordnung zu dem Perfusionsscore (z. Bsp. ps 2; 6 – 20% Perfusion) für das Medialstromgebiet ist hier mit höherer Reliabilität wahrscheinlich.

6.2.2 Interrater-Reliabilität der Perfusionsscores

Die statistische Auswertung der Interrater-Reliabilität für die beiden pro Serie vergebenen Perfusionsscores ergab zwischen *rater* und *senior rater* eine Übereinstimmung von 60% (*sliding-window* Übereinstimmung von 95%) und ein linear gewichtetes Cohens κ von 0.83, was einer sehr guten Übereinstimmung entspricht. Berücksichtigt man sowohl die feine

Graduierung der Perfusionsscores in insgesamt neun Klassen sowie den großen Erfahrungsunterschied zwischen *rater* und *senior rater* sind dies exzellente Ergebnisse, welche auch beim Vergleich mit den in Kapitel 2.3 beschriebenen Literaturwerten zur Interrater-Reliabilität eine potentielle Eignung als zum TICI Score alternative *ground truth* nahelegen. Ein wesentlicher Beitrag zu der hohen Übereinstimmung ist die Auflösung des Problems anatomischer Varianten sowie typischer Verschlusslokalisationen, welche eine eindeutige Zuordnung nach TICI erschweren. Weiter ist es naheliegend, dass ein Perfusionsscore perspektivisch besser mit dem klinischen *outcome* korreliert, da die Graduierung feiner ist und er anatomische Varianten, wie zum Beispiel eine stark ausgeprägte A. temporalis anterior, stärker abbildet. Für das Training neuronaler Netzwerke ist von enormem Vorteil, dass die Perfusionsscores bildmorphologisch eindeutig da unabhängig von der initialen Verschlusslokalisation sind. Eine aktuelle Studie von 2020, in der der neu eingeführte qTICI (quantitative TICI) vorgestellt wurde, zeigt das Interesse in Richtung Perfusion als besser objektivierbaren Parameter, wobei die Bestimmung des qTICI in dem Paper zu komplex für den klinischen Alltag ist.⁵⁰ Hier wäre es ausgesprochen attraktiv, die in Hypothese 1 und 2 untersuchten Ansätze zu fusionieren und mittels neuronaler Netzwerke Perfusionsprozente zu präzisieren.

7 Zusammenfassung

Im Rahmen dieser Arbeit sollte geklärt werden, ob eine automatisierte Beurteilung des Rekanalisationserfolgs nach TICI bei Schlaganfallpatienten anhand von zeitaufgelösten DSA-Serien, welche während der mechanischen Thrombektomie akquiriert wurden, durch neuronale Netzwerke möglich ist.

Zur Klärung dieser Frage wurden aus der Datenbank der Klinik für Neuroradiologische Diagnostik und Intervention des Universitätsklinikums Hamburg Eppendorf 491 DSA-Serien von 236 verschiedenen Patienten nach definierten Kriterien für einen Datensatz ausgewählt und bezüglich ihres TICI Scores reevaluiert. Mit Hilfe dieses Datensatzes wurden spezielle neuronale Netzwerke, Konvolutionsnetze, trainiert und ihre Prädiktionsgenauigkeit evaluiert. Zur Überprüfung der Robustheit des neuronalen Netzwerkes wurde ein externer, ungelabelter Datensatz an DSA Serien durch zwei rater zunächst unabhängig voneinander und im Falle eines abweichenden Scorings im Konsens geratet. Die Prädiktionsgenauigkeit des neuronalen Netzwerkes wurde auch auf den externen Datensatz bestimmt und statistisch ausgewertet.

Auf den *in-house* Testsatz ergab sich eine Übereinstimmung der Prädiktion zur angelegten *ground truth* von 71% bei einem Cohens κ von 0.61 und für den externen Datensatz eine Übereinstimmung von 74% und einem Cohens κ von 0.66. Bei repräsentativer Studiengröße ist diese Übereinstimmung im Bereich der in der Literatur beschriebenen Interrater-Variabilität durch erfahrene menschliche *rater*. Es konnte gezeigt werden, dass neuronale Netzwerke in der Lage sind, DSA Serien mechanischer Thrombektomien auf menschlichem Niveau nach TICI zu bewerten und somit einen wichtigen Beitrag zur besseren Vergleichbarkeit von Studien und im Bereich Qualitätsmanagement leisten können.

Daneben wurde ein zum TICI alternatives Scoring entwickelt und auf seine Interrater-Reliabilität untersucht. Bei einer *sliding-window* Übereinstimmung von 95% (absolut 60%) und einem gewichteten Cohens κ von 0.84 (ungewichtet κ 0.54) liegt die Interrater-Reliabilität trotz der feineren Graduierung des eingeführten Perfusionsscores in neun Klassen auf gleichem Niveau der in der Literatur beschriebenen Werte für den TICI Score. Die größere Robustheit des Scores gegenüber der Interpretation bei anatomischen Varianten und seiner bildmorphologischen Eindeutigkeit unabhängig von der initialen Verschlusslokalisation machen den Score gerade für das Training neuronaler Netzwerke zu einer interessanten Alternative zum TICI Score.

8 Summary

The aim of this study was to clarify whether an automated assessment of recanalization success according to the TICl score in stroke patients using time-resolved DSA series acquired during mechanical thrombectomy is possible.

To clarify this question, 491 DSA series from 236 different patients were selected from the database of the Department of Neuroradiological Diagnostics and Intervention of the University Medical Center Hamburg Eppendorf according to defined criteria for a data set and re-evaluated with respect to their assigned TICl score. Using this data set, specific neural networks, convolutional networks, were trained and their prediction accuracy was evaluated. To check the robustness of the neural network, an external unlabeled dataset of DSA series was initially rated by two experienced raters independently and in case of a discrepant scoring rerated in a consensus rating. The prediction accuracy of the neural network was also determined on the external dataset and statistically evaluated.

For the in-house test set, the agreement of the prediction to the applied ground truth was 71% with a Cohen's κ of 0.61. For the external data set, the agreement was 74% with a Cohen's κ of 0.66. For the underlying representative study size, this agreement is within the range of interrater variability by experienced human raters described in the literature. It could be shown that neural networks are able to evaluate DSA series of mechanical thrombectomies on a human level according to the TICl score and thus could make an important contribution to a better comparability of studies using the TICl score and in the field of quality management.

In addition, a scoring alternative to the TICl score based on the absolute perfusion of the respective hemisphere was developed and investigated for its interrater reliability. With a sliding-window accuracy of 95% (absolute 60%) and a weighted Cohen's κ of 0.84 (unweighted 0.54), the interrater reliability is on the same level as the values described in the literature for the TICl score, despite the finer graduation of the introduced perfusion score into nine distinct classes. The higher robustness of the score towards the interpretation of anatomical variants and its image morphological uniqueness independent of the initial occlusion location make the score an interesting alternative to the TICl score especially for the training of neural networks.

9 Literatur

- 1 Coupland A. P., Thapar A., Qureshi M. I., Jenkins H. and Davies A. H., The definition of stroke, *J R Soc Med*, **2017**, 110, 9-12.
- 2 Pearce J. M., Johann Jakob Wepfer (1620-95) and cerebral haemorrhage, *J Neurol Neurosurg Psychiatry*, **1997**, 62, 387.
- 3 National Institute of Neurological Disorders and Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke, *N Engl J Med*, **1995**, 333, 1581-1587.
- 4 Rother J., Ford G. A. and Thijs V. N., Thrombolytics in acute ischaemic stroke: historical perspective and future opportunities, *Cerebrovasc Dis*, **2013**, 35, 313-319.
- 5 Tillet W. S. and Garner R. L., The Fibrinolytic Activity of Hemolytic Streptococci, *J Exp Med*, **1933**, 58, 485-502.
- 6 Sussman B. J. and Fitch T. S., Thrombolysis with fibrinolytic in cerebral arterial occlusion, *J Am Med Assoc*, **1958**, 167, 1705-1709.
- 7 Campbell B. C., Meretoja A., Donnan G. A. and Davis S. M., Twenty-Year History of the Evolution of Stroke Thrombolysis With Intravenous Alteplase to Reduce Long-Term Disability, *Stroke*, **2015**, 46, 2341-2346.
- 8 Zeumer H., Hacke W. and Ringelstein E. B., Local intraarterial thrombolysis in vertebrobasilar thromboembolic disease, *AJNR Am J Neuroradiol*, **1983**, 4, 401-404.
- 9 Tsai F. Y., Berberian B., Matovich V., Lavin M. and Alfieri K., Percutaneous transluminal angioplasty adjunct to thrombolysis for acute middle cerebral artery rethrombosis, *AJNR Am J Neuroradiol*, **1994**, 15, 1823-1829.
- 10 Ueda T., Hatakeyama T., Kohno K., Kumon Y. and Sakaki S., Endovascular treatment for acute thrombotic occlusion of the middle cerebral artery: local intra-arterial thrombolysis combined with percutaneous transluminal angioplasty, *Neuroradiology*, **1997**, 39, 99-104.
- 11 Ueda T. et al., Angioplasty after intra-arterial thrombolysis for acute occlusion of intracranial arteries, *Stroke*, **1998**, 29, 2568-2574.
- 12 Ringer A. J., Qureshi A. I., Fessler R. D., Guterman L. R. and Hopkins L. N., Angioplasty of intracranial occlusion resistant to thrombolysis in acute ischemic stroke, *Neurosurgery*, **2001**, 48, 1282-1288; discussion 1288-1290.
- 13 Przybylowski C. J. et al., Evolution of endovascular mechanical thrombectomy for acute ischemic stroke, *World J Clin Cases*, **2014**, 2, 614-622.
- 14 Berkhemer O. A. et al., A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke, *New England Journal of Medicine*, **2015**, 372, 11-20.
- 15 Goyal M. et al., Randomized Assessment of Rapid Endovascular Treatment of Ischemic Stroke, *New England Journal of Medicine*, **2015**, 372, 1019-1030.
- 16 Saver J. L. et al., Stent-Retriever Thrombectomy after Intravenous t-PA vs. t-PA Alone in Stroke, *New England Journal of Medicine*, **2015**, 372, 2285-2295.
- 17 Jovin T. G. et al., Thrombectomy within 8 Hours after Symptom Onset in Ischemic Stroke, *New England Journal of Medicine*, **2015**, 372, 2296-2306.
- 18 Campbell B. C. et al., Endovascular therapy for ischemic stroke with perfusion-imaging selection, *N Engl J Med*, **2015**, 372, 1009-1018.
- 19 Powers W. J. et al., 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association, *Stroke*, **2018**, 49, e46-e110.
- 20 Ringleb P. A. and Veltkamp R., Ergänzung zur Leitlinie Schlaganfall Rekanalisierung, **2015**.

- 21 Turc G. *et al.*, European Stroke Organisation (ESO) - European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischaemic Stroke Endorsed by Stroke Alliance for Europe (SAFE), *Eur Stroke J*, **2019**, 4, 6-12.
- 22 Nogueira R. G. *et al.*, Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct, *N Engl J Med*, **2018**, 378, 11-21.
- 23 Hanahan D. and Weinberg R. A., The Hallmarks of Cancer, *Cell*, **2000**, 100, 57-70.
- 24 Higashida R. T. and Furlan A. J., Trial Design and Reporting Standards for Intra-Arterial Cerebral Thrombolysis for Acute Ischemic Stroke, *Stroke*, **2003**, 34, e109.
- 25 Kallmes D. F., TICI: if you are not confused, then you are not paying attention, *AJNR Am J Neuroradiol*, **2012**, 33, 975-976.
- 26 Zaidat O. O. *et al.*, Recommendations on Angiographic Revascularization Grading Standards for Acute Ischemic Stroke, *Stroke*, **2013**, 44, 2650.
- 27 Liebeskind D. S. *et al.*, eTICI reperfusion: defining success in endovascular stroke therapy, *J Neurointerv Surg*, **2019**, 11, 433-438.
- 28 Fugate J. E., Klunder A. M. and Kallmes D. F., What is meant by "TICI"?, *AJNR Am J Neuroradiol*, **2013**, 34, 1792-1797.
- 29 Noser E. A. *et al.*, Aggressive mechanical clot disruption: a safe adjunct to thrombolytic therapy in acute stroke?, *Stroke*, **2005**, 36, 292-296.
- 30 Goyal M. *et al.*, 2C or not 2C: defining an improved revascularization grading scale and the need for standardization of angiography outcomes in stroke trials, *J Neurointerv Surg*, **2014**, 6, 83-86.
- 31 Volny O., Cimflova P. and Szeder V., Inter-Rater Reliability for Thrombolysis in Cerebral Infarction with TICI 2c Category, *J Stroke Cerebrovasc Dis*, **2017**, 26, 992-994.
- 32 Suh S. H. *et al.*, Clarifying differences among thrombolysis in cerebral infarction scale variants: is the artery half open or half closed?, *Stroke*, **2013**, 44, 1166-1168.
- 33 Bar M. *et al.*, Diagnosis of recanalization of the intracranial artery has poor inter-rater reliability, *AJNR Am J Neuroradiol*, **2012**, 33, 972-974.
- 34 Zhang G. *et al.*, Operator Versus Core Lab Adjudication of Reperfusion After Endovascular Treatment of Acute Ischemic Stroke, *Stroke*, **2018**, 49, 2376-2382.
- 35 Zaharchuk G., Gong E., Wintermark M., Rubin D. and Langlotz C. P., Deep Learning in Neuroradiology, *American Journal of Neuroradiology*, **2018**.
- 36 LeCun Y., Bengio Y. and Hinton G., Deep learning, *Nature*, **2015**, 521, 436.
- 37 Krizhevsky A., Sutskever I. and Hinton G. E., ImageNet Classification with Deep Convolutional Neural Networks, **2012**, 1097--1105.
- 38 Yamashita R., Nishio M., Do R. K. G. and Togashi K., Convolutional neural networks: an overview and application in radiology, *Insights into Imaging*, **2018**.
- 39 Gulshan V. *et al.*, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, **2016**, 316, 2402-2410.
- 40 Esteva A. *et al.*, Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, **2017**, 542, 115.
- 41 Hekler A. *et al.*, Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images, *Eur J Cancer*, **2019**, 118, 91-96.
- 42 Litjens G. *et al.*, A Survey on Deep Learning in Medical Image Analysis, *arXiv:1702.05747*, **2017**.

- 43 Siemens-Healthcare. *Artificial intelligence - Transforming data into knowledge for better care*, <https://www.healthcare.siemens.com/infrastructure-it/artificial-intelligence>
- 44 Hosny A., Parmar C., Quackenbush J., Schwartz L. H. and Aerts H. J. W. L., Artificial intelligence in radiology, *Nature Reviews Cancer*, **2018**, 18, 500-510.
- 45 Goyal M. *et al.*,. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials, *The Lancet*, **2016**, 387, 1723-1731.
- 46 Busch M. A. and Kuhnert R. Vol. 2 (Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung, 2017).
- 47 Wareham J. *et al.*,. Factors impacting on technical success in stroke thrombectomy: experience of a UK neuro-interventional unit, *Clinical Radiology*, **2019**, 74, 390-398.
- 48 Rinaldo L., Brinjikji W. and Rabinstein A. A., Transfer to High-Volume Centers Associated With Reduced Mortality After Endovascular Treatment of Acute Stroke, *Stroke*, **2017**, 48, 1316-1321.
- 49 Zarrinkoob L. *et al.*,. Blood flow distribution in cerebral arteries, *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, **2015**, 35, 648-654.
- 50 Prasetya H. *et al.*,. qTICI: Quantitative assessment of brain tissue reperfusion on digital subtraction angiograms of acute ischemic stroke patients, *Int J Stroke*, **2020**, 1747493020909632.

10 Erklärung des Eigenanteils an der Publikation

- Erhebung des *in-house* Datensatzes
 - Sichtung und Dokumentation aller Serien mit Verschluss des vorderen Kreislaufs
 - Entwicklung des methodischen Vorgehens bei Erhebung des Datensatzes
 - Extraktion des TICI Scores aus dem Befund und Plausibilitätsprüfung
 - Rating mit ggf. anschließendem Konsensrating von nicht aus dem Befund extrahierbaren Scores
- Erhebung des externen Datensatzes
 - Sichtung und Bewertung der zur Verfügung stehenden Serien
 - Zweizeitiges Scoring der inkludierten Serien und Konsensrating zusammen mit Dr. med. Fabian Flottmann
- Auswertung und Interpretation der prädizierten Scores auf die beiden Datensätze
 - Statistische Analyse
 - Bildmorphologische Evaluation der falsch prädizierten Scores
- Schreiben der Publikation zusammen mit den Co-Autoren
 - Anfertigung des Manuskripts
 - Literaturrecherche
 - Anfertigung der veröffentlichten Videosequenzen

11 Danksagung

Ich möchte Herrn Prof. Jens Fiehler für die Überlassung des wirklich interessanten Themas sowie Andreas Frölich und Fabian Flottmann für die Betreuung und Unterstützung bei der Durchführung der Arbeit. Das nötige Vertrauen, die fachliche Unterstützung und das ausgesprochen gute und hilfsbereite Miteinander haben nicht nur zum Gelingen dieser Arbeit beigetragen, sondern auch dazu, dass mir die Arbeit an dieser unheimlich viel Freude gemacht hat.

Mein besonderer Dank gilt unseren Kooperationspartnern René Werner, Maximilian Nielsen und Thilo Sentker aus dem *Institute for Computational Neurosciences*, ohne die die Beantwortung der Fragestellung überhaupt nicht möglich gewesen wäre. Auch hier war die Zusammenarbeit ausgesprochen gut und freundschaftlich und hat mir viel Freude bereitet.

Weiter danke ich allen Kollegen aus der Klinik und Poliklinik für neuroradiologische Diagnostik und Intervention für die gute Zeit, die ich hier hatte, für die freundliche Aufnahme ins Team und für das Verständnis, wenn ich mal wieder einen Befundungsrechner stundenlang in Anspruch genommen habe.

Zu guter Letzt möchte ich meiner Frau danken, die mir völlig selbstverständlich die Zeit gegeben hat, um diese Arbeit überhaupt anfertigen zu können, in dem sie sich mit viel Liebe und Engelsgeduld um unsere Kinder gekümmert hat. Ich bewundere dich dafür, wie du das alles schaffst!

12 Lebenslauf

Der Lebenslauf wurde aus datenschutzrechtlichen Gründen entfernt.

13 Eidesstattliche Versicherung

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: