

Aus dem CharitéCentrum für Neurologie, Neurochirurgie und Psychiatrie
CC15 Klinik für Neurochirurgie
Direktor: Prof. Dr. Peter Vajkoczy

Habilitationsschrift

KI-basierte Personalisierung der Therapie des Schlaganfalls zur Outcome-Verbesserung

zur Erlangung der Lehrbefähigung für das Fach Neurochirurgie
vorgelegt dem Fakultätsrat der Medizinischen Fakultät Charité-Universitätsmedizin Berlin

von

Dr. med. Dietmar Frey

Eingereicht: 06/2023
Dekan: Prof. Dr. med. Joachim Spranger

1. Gutachter/in:
2. Gutachter/in:

1. EINLEITUNG	3
2. EIGENE ARBEITEN	6
2.1 Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome	6
2.2 A precision medicine framework for personalized simulation of hemodynamics in cerebrovascular disease	22
2.3 Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks	43
2.4 Toward Sharing Brain Images: Differentially Private TOF-MRA Images With Segmentation Labels Using Generative Adversarial Networks	53
2.5 Comparing poor and favorable outcome prediction with machine learning after mechanical thrombectomy in acute ischemic stroke	66
3. DISKUSSION	78
4. ZUSAMMENFASSUNG	81
5. LITERATURANGABEN	82
6. ABKÜRZUNGSVERZEICHNIS	87
7. DANKSAGUNG	88
8. ERKLÄRUNG	89

1. Einleitung

Weltweit ist der Schlaganfall eine der Hauptursachen für Tod und Behinderung¹. Ohne wirksamere Prävention, Diagnostik und Therapie werden Leiden der Patienten und ihrer Angehörigen sowie die ökonomische Belastung für die Gesellschaft weiter zunehmen¹. Die Entscheidung über die Therapie bei Patienten mit akutem ischämischem Schlaganfall basiert auf prospektiven und randomisierten Studien mit relativ hohen Fallzahlen, in denen sich die Zeitdauer zwischen Symptombeginn, Behandlungszeitpunkt und Veränderungen in der MRT-Diagnostik als wichtige Parameter und Biomarker herausgestellt haben². Der Nachteil dieser Strategien ist allerdings, dass hierfür populationsbasierte Schwellenwerte verwendet werden, die die individuellen Merkmale der jeweiligen Patienten nur unzureichend berücksichtigen^{2,3,4}. Mit der zunehmenden Entwicklung von Methoden der künstlichen Intelligenz (KI) - hier insbesondere Machine und Deep Learning-basierte Ansätze - wird angestrebt, die potentiell entscheidenden Informationen, die in großen Datenmengen liegen, für eine bessere Prädiktion verschiedener Therapieoptionen nutzbar zu machen. Dieser datengetriebene und KI-basierte Ansatz kann die Schlaganfallversorgung durch den Einsatz intelligenter und individualisierter Entscheidungshilfen revolutionieren.

Entscheidungshilfe und -unterstützung

Methoden der Künstlichen Intelligenz (KI) können als Grundlage für Algorithmus-basierte Entscheidungshilfen dienen, die den Ärzten, die über die Therapie entscheiden, zusätzliche Informationen oder Hinweise geben⁵. Für verschiedene Schritte in der Diagnostik wurden für den Schlaganfall bereits KI-basierte Methoden zur Entscheidungshilfe entwickelt. Für die Berechnung des Alberta Stroke Program Early CT Score (ASPECTS) wurde ein Software-Werkzeug entwickelt^{6,7,8,9}. Des Weiteren wurden Automatisierungssequenzen für die Identifizierung von Biomarker für ischämischer Läsionen entwickelt^{10,11,12}. Soweit veröffentlicht, existieren allerdings noch keine Algorithmen-basierten, individualisierten und personalisierten Lösungen für eine integrative Entscheidung, welche Behandlung den besten Erfolg im Individualfall für den konkreten Patienten bewirkt.

Im klinischen Alltag werden zunehmend Lösungen zur Entscheidungsunterstützung eingesetzt, um effizientere Arbeitsabläufe zu fördern und das Outcome der Patienten zu verbessern⁵. Vorausgesetzt, dass diese Lösungen validiert sind, das heißt in klinischen Studien durch den Einsatz KI-basierter Decision Support Systems ein messbarer Vorteil (Benefit) für den Patienten entsteht, werden technologie-basierte innovative Lösungen für Patienten und Ärzte von großem Nutzen sein.

In anderen Bereichen wie in der Entwicklung für das autonome Fahren und im Feld Computer Vision werden Techniken des Machine Learning (ML) für die prädiktive Modellierung vielfach erfolgreich eingesetzt und reflektieren den aktuellen Stand der Wissenschaft und Technik¹³. Für die Anwendung im medizinischen Kontext ist der Einsatz als klinische Entscheidungshilfe aufgrund des oft fehlenden klinischen Nachweises der Überlegenheit der Methode zum Goldstandard (Benefit) sowie der - auch damit verbundenen - fehlenden Akzeptanz der Ärztinnen und Ärzte und der regulatorischen Anforderungen noch in der Anfangsphase. Es werden momentan Methoden vor allem dann eingesetzt, um klinische Werte zur Vorhersage einer Diagnose, eines Ergebnisses oder eines Risikos zu gewinnen^{14,15}.

Machine und Deep Learning

Künstliche Intelligenz (KI), insbesondere Machine Learning (ML), ist zu einer der wichtigsten Technologien geworden, die die sogenannte vierte industrielle Revolution vorantreiben¹⁶. Die derzeit vielversprechendsten ML-Methoden sind tiefe künstliche neuronale Netze (artificial neural nets = ANN). Vereinfacht gesagt, sind ANNs neuronalen Strukturen nachempfunden und bestehen aus mehreren Schichten künstlicher Neuronen¹⁷. Damit wird die Stärke der Verbindungen zwischen den Neuronen verschiedener Schichten im Trainingsprozess berechnet und fließt in die Vorhersagefähigkeit des Modells ein. ANNs sind allgemeine Funktionsapproximatoren, die theoretisch in der Lage sind, jede mathematische Funktion von Daten zu approximieren¹⁸. In den letzten Jahren werden Ansätze des Machine Learning und Deep Learning zunehmend für den Einsatz in der Medizin entwickelt. Zu diesen Techniken gehören Support-Vektor-Maschinen (SVM), Entscheidungsbäume (Decision Trees), Bayes'sche Ansätze und künstliche neuronale Netze (ANN). Mit diesen Methoden kann die klinische Leistung von Prognosemodellen verbessert werden¹⁷. Im Speziellen zeigen künstliche neuronale Netze (ANN), sogenannte Ensemblemodelle und Methoden des Tree Boosting - ein auf Entscheidungsbäumen basierender Algorithmus – im Vergleich zu traditionellen ML-Ansätzen wie die lineare und logistische Regression eine überlegene Genauigkeit und Performanz^{19,20,21,22,23,24,25}. Allerdings kann die Heterogenität der Parameter, der Datenquellen und der klinischen Fragestellungen sowie die Vielzahl der angewandten Algorithmen die Nachvollziehbarkeit und Reproduzierbarkeit der Methode einschränken²¹.

Methoden der Künstlichen Intelligenz können primär dazu verwendet werden, KI-basierte klinische Entscheidungshilfen bereitzustellen, die den behandelnden Ärzten zusätzliche Informationen oder Hinweise liefern⁵. Diese klinischen Entscheidungshilfesysteme (clinical decision support systems, CDSS) wurden bereits für relativ einfache Fragestellungen und Anwendungsfälle entwickelt und sind kommerziell erhältlich. Für komplexere Entscheidungen wie eine multidimensionale und integrative Therapiestratifizierung oder eine Outcome-Prädiktion stehen allerdings noch keine Algorithmus-basierten, individualisierten Lösungen zur Verfügung. Es wird erwartet, dass klinische Entscheidungsunterstützungssysteme durch Personalisierung der Therapie zu besseren Behandlungsergebnissen und effizienteren Arbeitsabläufen in der klinischen Praxis führen können und daher sowohl für Patienten als auch für die behandelnden Ärzte ein großes Potenzial bieten.

Das Black Box Dilemma

Eine Kritik an innovativen Machine und Deep Learning-basierten Ansätzen und Technologien ist, dass Nachvollziehbarkeit, Erklärbarkeit und Transparenz der Modelle nicht vollständig gegeben sei. Mit anderen Worten, dass die generierten Vorhersagemodelle nicht erklärt werden können und dies ein Hindernis auf Akzeptanz und Implementierung in der Klinik darstellen könne. In der inhärenten Charakteristik der Ansätze des Machine und insbesondere Deep Learning liegt es, dass die multiplen Wiederholungen/Iterationen der Trainings- und Testing-Läufe in verborgene Ebenen („hidden layers“) erfolgen und sich damit einer vollständigen Erklärbarkeit und Nachvollziehbarkeit entziehen. Ein absoluter Anspruch an Nachvollziehbarkeit und Transparenz kann allerdings bedeuten, dass eine

Erhöhung der Performanz von Modellen nicht in deren unmittelbare Verwertung und Anwendung einfließen kann²⁵. Im Gegensatz zu den hoch performanten Methoden bieten traditionellere und klassische Ansätze leichtere Interpretationen der generierten Ergebnisse an und können leichter in die klinische Anwendung überführt werden²⁶.

Grundsätzlich gilt, dass eine sinnvolle Balance zwischen dem Bedürfnis nach Erklärbarkeit und Transparenz und einer signifikanten Erhöhung der Leistungsfähigkeit von neuen innovativen Modellen gefunden werden muss. Im medizinischen Feld ist in Abgrenzung zu anderen Industrien ein erhöhtes Maß an Interpretierbarkeit und Erklärbarkeit erforderlich. Während der Entwicklung muss daher erstens eine Interpretation und sichere Überprüfung der gewonnenen Ergebnisse möglich sein²⁷. Im Hinblick auf Bias (Verzerrungen/Vorurteile in den Daten) muss zweitens eine Bewertung der Sicherheit und Fairness medizinischer Produkte gegeben sein. Dies ist auch im Hinblick auf eine mögliche Zertifizierung als Medizinprodukt erforderlich²⁸. Schließlich müssen Akzeptanz und das Vertrauen medizinischen Fachkräfte als Anwender der klinischen Unterstützungssysteme berücksichtigt werden.

Teilweise wird die Ansicht vertreten, dass eine Anwendung dieser Ansätze aufgrund der nicht nachvollziehbaren Black Box nicht mit den genannten Anforderungen in Einklang gebracht werden kann^{29,30}. Dies kann für Forscher und Entwickler zu folgendem Dilemma führen: Entweder sie verwenden Methoden mit potenziell höherer Leistung, aber Intransparenz. Oder sie verwenden niedrig performante Methoden, die allerdings eine Erklärbarkeit bieten, um die ethischen und gesetzlichen Anforderungen zu erfüllen²⁵.

Aus diesen Gründen wurden in den vergangenen Jahren Methoden für die Interpretierbarkeit und Erklärbarkeit (explainable artificial intelligence/xAI) entwickelt, die auch auf innovative und leistungsstarke Modelle und Algorithmen angewandt werden können. So wurde ein Ansatz vorgestellt für die Erklärbarkeit künstlicher neuronaler Netze (ANN), für das Tree Boosting kann die Berechnung für die Gewichtung der Input Features (Weighing) und eine Rangfolge der Input Features (Ranking) erfolgen³¹. Dies kann ein Weg sein, um die Transparenz von KI-basierten Systemen zu erhöhen.

Insgesamt soll mit dieser Arbeit gezeigt werden, dass die Entwicklung und der Einsatz von innovativen Methoden aus der Künstlichen Intelligenz das Potential haben, durch Analyse und Integration multidimensionaler Daten das Outcome für den individuellen Patienten oder die individuelle Patientin in der Schlaganfallbehandlung zu verbessern. Darüber hinaus wird die Erklärbarkeit von KI-Modellen als wesentliches Element für Akzeptanz und Implementation in den Fokus gebracht. Damit soll die Grundlage für klinisch wirksame und ethische Entscheidungsunterstützungssysteme gelegt werden, um eine das Patientenwohl fördernde digitale personalisierte Medizin zu ermöglichen.

2. Eigene Arbeiten

Die folgenden fünf inhaltlich kohärenten Arbeiten werden mit Überleitungstexten in Zusammenhang gebracht. Zunächst wird die initiale Einführungsarbeit zu klinischen Entscheidungsunterstützungssystemen in der Schlaganfalltherapie vorgestellt.

2.1 Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome³²

Moderne Methoden des maschinellen Lernens (ML) und der künstlichen Intelligenz (KI) werden zunehmend bei der prädiktiven Modellierung des klinischen Outcomes eingesetzt, um Ärzten klinische Entscheidungshilfesysteme zur Verfügung zu stellen. Hierbei schneiden moderne ML-Ansätze wie künstliche neuronale Netze (ANNs) und Tree Boosting oft besser ab als traditionellere Methoden wie die logistische Regression. Der Nachteil dieser moderneren Methoden wie ANNs oder Tree Boosting ist, dass eine Nachverfolgbarkeit des KI-Prozesses und damit eine Transparenz des Algorithmus nur eingeschränkt möglich erscheint. Mit anderen Worten ist eine Erklärbarkeit wie die modellierten Vorhersagen zustande kommen, nicht vollständig möglich. Im medizinischen Bereich ist diese Erklärbarkeit und Transparenz der angewandten Modelle jedoch von entscheidender Bedeutung, insbesondere wenn diese Algorithmen als Basis für klinische Entscheidungsunterstützungssysteme dienen. In den letzten Jahren wurden aus diesem Grund vielfältige Interpretationsmethoden für moderne ML-Methoden entwickelt.

In der hier vorgelegten Arbeit stellen wir den ersten Vergleich der Erklärbarkeit von zwei modernen ML-Methoden - Tree Boosting und Multilayer Perceptrons (MLPs) - mit traditionellen logistischen Regressionsmethoden anhand eines Paradigmas zur Vorhersage von Schlaganfällen vor. Hierzu wurden klinische Merkmale zur Vorhersage eines dichotomisierten Modified Rankin Scale (mRS) Scores 90 Tage nach dem Schlaganfall verwendet.

Um eine Interpretierbarkeit, bzw. Erklärbarkeit zu ermöglichen, bewerteten wir die Bedeutung und das jeweilige Gewicht der klinischen Merkmale für die Vorhersage mit Hilfe von Deep Taylor Decomposition für Multilayer Perceptrons (MLPs), Shapley-Werten für Tree Boosting und Modellkoeffizienten für logistische Regression.

In Bezug auf die Leistung, die anhand der Werte für die AUC (Area under Curve) im Testdatensatz gemessen wurde, wurden für alle Modelle vergleichbare Ergebnisse erzielt: Die AUC-Werte der logistischen Regression lagen bei 0,83, 0,83 und 0,81 für drei verschiedene Regularisierungsverfahren; die AUC des Tree Boosting betrug 0,81; die AUC des MLP lag bei 0,83.

Darüber hinaus ergab die Analyse der Interpretierbarkeit konsistente Ergebnisse für alle Modelle, in dem das Alter der Patienten und der Schweregrad des Schlaganfalls als die wichtigsten prädiktiven Merkmale eingestuft wurden. Bei weniger wichtigen Merkmalen wurden einige Unterschiede zwischen den Methoden festgestellt. Diese Ergebnisse zeigen, dass moderne Methoden des maschinellen Lernens eine Erklärbarkeit bieten können, die in Übereinstimmung mit Expertenwissen steht.

RESEARCH ARTICLE

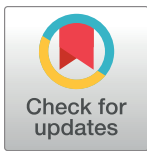
Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome

Esra Zihni¹ , Vince Istvan Madai¹ *, Michelle Livne¹ , Ivana Galinovic² , Ahmed A. Khalil² , Jochen B. Fiebach² , Dietmar Frey¹ 

1 Charité Lab for Artificial Intelligence in Medicine—CLAIM, Charité - Universitätsmedizin Berlin, Berlin, Germany, **2** Centre for Stroke Research Berlin, Charité - Universitätsmedizin Berlin, Berlin, Germany

 These authors contributed equally to this work.

* vince_istvan.madai@charite.de



Abstract

State-of-the-art machine learning (ML) artificial intelligence methods are increasingly leveraged in clinical predictive modeling to provide clinical decision support systems to physicians. Modern ML approaches such as artificial neural networks (ANNs) and tree boosting often perform better than more traditional methods like logistic regression. On the other hand, these modern methods yield a limited understanding of the resulting predictions. However, in the medical domain, understanding of applied models is essential, in particular, when informing clinical decision support. Thus, in recent years, interpretability methods for modern ML methods have emerged to potentially allow explainable predictions paired with high performance. To our knowledge, we present in this work the first explainability comparison of two modern ML methods, tree boosting and multilayer perceptrons (MLPs), to traditional logistic regression methods using a stroke outcome prediction paradigm. Here, we used clinical features to predict a dichotomized 90 days post-stroke modified Rankin Scale (mRS) score. For interpretability, we evaluated clinical features' importance with regard to predictions using deep Taylor decomposition for MLP, Shapley values for tree boosting and model coefficients for logistic regression. With regard to performance as measured by Area under the Curve (AUC) values on the test dataset, all models performed comparably: Logistic regression AUCs were 0.83, 0.83, 0.81 for three different regularization schemes; tree boosting AUC was 0.81; MLP AUC was 0.83. Importantly, the interpretability analysis demonstrated consistent results across models by rating age and stroke severity consecutively amongst the most important predictive features. For less important features, some differences were observed between the methods. Our analysis suggests that modern machine learning methods can provide explainability which is compatible with domain knowledge interpretation and traditional method rankings. Future work should focus on replication of these findings in other datasets and further testing of different explainability methods.

OPEN ACCESS

Citation: Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. (2020) Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. PLoS ONE 15(4): e0231166. <https://doi.org/10.1371/journal.pone.0231166>

Editor: Ruxandra Stoean, University of Craiova, ROMANIA

Received: November 8, 2019

Accepted: March 17, 2020

Published: April 6, 2020

Copyright: © 2020 Zihni et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data cannot be shared publicly because of data protection laws imposed by institutional ethics committee guidelines. Data might be available from the institutional ethics committee of Charité Universitätsmedizin Berlin (contact via ethikkommission@charite.de) for researchers who meet the criteria for access to confidential data.

Funding: This work has received funding by the German Federal Ministry of Education and Research through (1) the grant Centre for Stroke

Research Berlin and (2) a Go-Bio grant for the research group PREDICTioN2020 (lead: DF). No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Machine learning (ML) techniques are state-of-the-art in predictive modeling in fields like computer vision and autonomous navigation [1]. Increasingly, these tools are leveraged for clinical predictive modeling and clinical decision support, where clinical values are used to predict a clinical status, e.g. a diagnosis, outcome or risk [2,3]. Here, newer machine learning techniques—we will refer to them as modern machine learning techniques in this work—including artificial neural nets (ANN), especially deep learning (DL), and ensemble models such as tree boosting have often shown higher performance than traditional machine learning techniques such as linear or logistic regression, e.g. [4–8].

However, a common criticism of these modern techniques is that while they might increase model performance they do not provide the possibility to explain the resulting predictions [9]. In contrast, traditional techniques allow explanations by various means and this approach has been the backbone of explainable clinical predictive modeling to date [10]. The necessity of interpretable ML systems are of particular concern in the medical domain. An explainable AI system is essential to provide: 1) Interpretation and safe check of the acquired results during development [11]. 2) Better assessment of safety and fairness of medical products, especially regarding bias, during the regulatory process [12]. 3) Domain knowledge supported interpretation leading to increased trust by the physicians, other healthcare professionals, and patients [12]: Some argue that black box approaches are unacceptable for clinical decision support from the physician's point-of-view [13] and from the patient's point-of-view [14]. Thus, currently, researchers and developers are facing an unfortunate trade-off: either to use methods with potentially higher performance or to use methods providing explainability to comply with ethical and regulatory requirements [9].

Fortunately, interpretability methods tailored to modern machine learning algorithms have emerged lately, therefore potentially allowing high performance and explainable models. For one, in the last few years several techniques have been developed to open the most notorious black box, namely artificial neural networks and provide explainable models [11]. Moreover, tree boosting provides high performance clinical predictive modeling and also allows the calculation of feature importance and ranking, e.g. Lundberg et al [15]. However, to our knowledge, these approaches have not yet been compared to the traditional methods in terms of interpretability for clinical predictive modeling.

In the present work, we thus compared the above mentioned two modern ML methods, ANNs and tree boosting, to traditional methods with regard to explainability. We chose a well-characterized stroke clinical outcome paradigm. Here, available clinical features such as age, the severity of the stroke or information about treatment are used to predict the 3 months post-stroke outcome. Many replications in the past have established main factors driving the prediction, namely age and stroke severity, e.g. [16–19]. Thus, within this paradigm, modern machine learning explanations can be interpreted against a baseline. Concretely, we used a multilayer perceptron (MLP) with deep Taylor decomposition as an example for an explainable ANN approach [20], the CATBOOST algorithm with Shapley Additive exPlanations (SHAP) values as an example for explainable tree boosting [15] and compared performance and explainability with different versions of (regularized) logistic regression for a binary outcome (GLM, LASSO, and Elastic Net).

Methods

Patients and clinical metadata pre-processing

In a retrospective analysis, patients with acute ischemic stroke from the 1000plus study were included [21]. The study was approved by the institutional ethics committee of Charité

Universitätsmedizin Berlin in accordance with the Helsinki declaration and all patients gave written informed consent. Patients were triaged into receiving iv-tissue-plasminogen-activator (tPA) for thrombolysis therapy or conservative therapy. The modified Rankin Scale (mRS), representing the degree of disability or dependence in the daily activities, was assessed for each patient 3 months post-stroke via a telephone call. The available database consisted of 514 patients who received imaging at 3 imaging time points. Of these, 104 were lost-to-follow-up and had no mRS values. 1 patient had to be excluded due to values outside of the possible parameter range. Moreover, 95 patients had to be excluded due to infratentorial stroke and no visible diffusion-weighted imaging (DWI) lesions. Specific further inclusion criteria of our sub-study were a ratio of at least 1 to 4 for binary variables (absence/presence) and no more than 5% missing values resulting in the final number of 314 patients and the following clinical parameters for the predictive models: age, sex, initial NIHSS (National Institute of Health Stroke Scale; measuring stroke severity), history of cardiac disease, history of diabetes mellitus, presence of hypercholesterolemia, and thrombolysis treatment. For a summary of the patients' clinical features and their distribution, see [Table 1](#).

Data accessibility

Data cannot be shared publicly because of data protection laws imposed by institutional ethics committee guidelines. Data might be available from the institutional ethics committee of Charité Universitätsmedizin Berlin (contact via ethikkommission@charite.de) for researchers who meet the criteria for access to confidential data. The code used in the manuscript is available on Github (<https://github.com/prediction2020/explainable-predictive-models>).

Outcome prediction supervised machine learning framework

In a supervised machine-learning framework, the clinical parameters ([Table 1](#)) were used to predict the final outcome of stroke patients in terms of dichotomized 3-months post-stroke mRS, where $mRS \in \{0,1,2\}$ indicates a good outcome (i.e. class label for a given observation i) and $mRS \in \{3,4,5,6\}$ indicates a bad outcome (i.e. class label for a given observation i). The applied dichotomization resulted in 88 positive (i.e. bad outcome) and 226 negative (i.e. good outcome) classes.

Feature multicollinearity

Importantly, methods for feature ranking can be influenced by feature multicollinearity. Particularly, Beta weights in regression analysis can be erroneous in case of multicollinearity

Table 1. Summary of the clinical data.

Clinical information	Value
Median age (IQR)	72 (15)
Sex (Females/ Males)	196 / 118
Median initial NIHSS (IQR)	3 (5)
Cardiac history (yes/ no)	84 / 230
Diabetes mellitus (yes/ no)	79 / 235
Hypercholesterolemia (yes/ no)	182 / 132
Thrombolysis (yes / no)	74 / 240

The table summarizes the distribution of the selected clinical data covariates acquired in the acute clinical setting. NIHSS stands for National Institutes of Health Stroke Scale; IQR indicates the interquartile range.

<https://doi.org/10.1371/journal.pone.0231166.t001>

[22,23] and certain applications of feature importance calculation for tree boosting are simplified under the assumption of feature independence. To ensure an unbiased comparison of the models interpretability we estimated multicollinearity of the features using the variance inflation factor (VIF) [24]. The chosen features in the analysis demonstrated negligible multicollinearity with VIFs < 1.91 (Age: 1.15; Sex: 1.91, NIHSS: 1.28; Cardiac history: 1.33; Diabetes: 1.36; Hypercholesterolemia: 1.74; Thrombolysis: 1.50). This makes our stroke outcome paradigm particularly suited to compare explainability.

Predictive modeling and Interpretability

In this study, machine-learning (ML) methods were applied to predict the final outcome based on clinical data. In the context of tabular data as in the given study, the interpretability of the resulting models corresponds to a rating of feature importance. The interpretability frameworks suggested in this study are tailored to the models and therefore indicate the relative contribution of the features to the respective model prediction. The different ML algorithms and the corresponding interpretability derivations are described as follows.

Traditional (linear) ML frameworks. 1. *Generalized Linear Model (GLM)*. GLM is a generalization of linear regression that allows for a response to be dichotomous instead of continuous. Hence the model predicts the probability of a bad outcome (vs. good outcome) based on a set of explanatory variables according to the following relation:

$$P(O = 1|\bar{X}) = \frac{1}{1 + e^{-\sum_i \beta_i x_i}}$$

where $P(O = 1|\bar{X})$ is the probability for a bad outcome ($O = 1$) given the vector of corresponding covariates \bar{X} .

β stands for model parameterization. The objective function for the optimization problem is defined by maximum likelihood estimation (MLE):

$$J(\bar{\beta}) = \ln \prod_{i=1}^N P(O_i = 1|\bar{X}_i, \bar{\beta})$$

where $J(\bar{\beta})$ stands for the objective function for the given model parametrization, $P(O_i = 1|\bar{X}_i, \bar{\beta})$ is the predicted outcome probability for the given covariates \bar{X}_i and model parametrization $\bar{\beta}$ and N is the number of observations. In this formulation, this special case of a GLM is also known as logistic regression.

2. *Lasso*. Lasso, standing for least absolute shrinkage and selection operator, provides the L1 regularized version of GLM. An L1 penalization of the model parametrization reduces overfitting of the model and is applied by the addition of the L1 regularization term to the objective function:

$$J_L(\bar{\beta}) = J(\bar{\beta}) + \alpha \|\bar{\beta}\|$$

where $J_L(\bar{\beta})$ stands for the Lasso objective function and α is the scaling factor hyperparameter.

3. *Elastic Nets*. Similarly to Lasso, elastic net provide a regularized variate of the GLM. Here two types of regularization terms are added to the objective function that provide L1 and L2 penalization of the model parametrization respectively:

$$J_{EN}(\bar{\beta}) = J(\bar{\beta}) + \alpha \|\bar{\beta}\| + \gamma \|\bar{\beta}^2\|$$

where $J_{EN}(\bar{\beta})$ stands for the elastic nets objective function and α and γ are the scaling factors hyperparameters.

For the three linear models, the interpretability of the models was deduced using the resulted model parametrization. Hence, the rating of the features was derived by the values of the model coefficients β . As outlined above, this is sufficient since our features do not exhibit collinearity [23].

Modern (nonlinear) ML frameworks. 4. *Tree boosting (CatBoost)*. Tree boosting solves the described classification problem by producing a prediction model as an ensemble of weak classification models, i.e. classifiers. As an ensemble method, the algorithm builds many weak classifiers in the form of decision trees and then integrates them into one cumulative prediction model to obtain better performance than any of the constituent classifiers. The prediction is then given using K additive functions:

$$P(O = 1|\bar{X}) = \sum_{k=1}^K f_k(\bar{X}), f_k \in \mathcal{F}$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\} (q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ is the space of regression trees. Here q denotes the structure of each tree and T is the number of leaves in the tree. Each $f(x)$ represents an independent tree structure q and leaf weights w . The output of the regression trees is a continuous score represented by w_i for leaf i . Each observation is classified using each constituent tree to the corresponding leafs and the outcome prediction $P(O = 1|\bar{X})$ is finally calculated as the cumulative sum of scores of the corresponding leafs. The objective function for optimization constitutes of the convex loss function, here chosen as logistic function, and a regularization component:

$$J_c(\varphi) = \sum_i l(y'_i, y_i) + \sum_k \Omega(f_k)$$

where the convex loss is given by:

$$l(y'_i = P(O = 1|\bar{X}), y_i) = \frac{-\sum_{i=1}^N w_i (y_i \log(y'_i) + (1 - y_i) \log(1 - y'_i))}{\sum_{i=1}^N w_i}$$

which is the logistic loss and the regularization component is given by:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

where w are the model weights penalized through L2 normalization and T is again the number of leaves in the tree. Here φ represents the corresponding model parametrization. In this study we used the CATBOOST module to implement the tree boosting model allowing to successfully integrate both numerical and categorical features [25].

In the context of tree boosting models, SHapley Additive exPlanations (SHAP) values construct a robust unified interpretability framework, breaking down the prediction to show the impact of each input feature [15]. The SHAP values attribute to each feature the average change in the model prediction when that feature is integrated to the model. It calculates a marginal contribution of the feature by averaging over every possible sequence in which that feature could have been introduced to make the prediction. This allows for calculating the contribution of the feature to the final decision irrespective of in which order it was used in the decision tree. The Shapley value of an input feature i for a single observation is calculated as

follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

where F is the set of all input features, $|F|$ representing its size. S represents any subset of input features that was introduced to the model before feature i , and $|S|$ is the size of that subset. The second factorial in the nominator then gives the size of the remaining subset of input features that will succeed feature i . The final multiplicative factor quantifies the difference in the model prediction when feature i is introduced.

Finally, the overall rating of the feature contribution to the model is then achieved by averaging the SHAP values over all observations.

5. *MLP*. A multilayer perceptron (MLP) is a type of feedforward artificial neural network that is composed of connectionist neurons, also known as perceptrons, in a layered structure. An MLP architecture is constructed of 3 components: 1) an input layer to receive the information 2) an output layer that makes a decision or prediction about the input and 3) one or more hidden layers that allow for feature extraction and modeling of the covariates dynamics using nonlinear transformations. According to the universal approximation theorem, an MLP with one hidden layer can approximate any function [26].

Here the model prediction is given by:

$$P(O = 1|\bar{X}) = f(g(a(g(\bar{X}))))$$

where $f(x_k) = \frac{\exp(x_k)}{\sum_c \exp(x_c)}$ is the (softmax) output layer activation, k is the predicted class and c is any of the possible classes for prediction. a denotes the hidden layer activation function where M represents the number of nodes in the layer.

The core objective function utilized for the MLP model was binary cross-entropy:

$$J_m(\varphi) = -\frac{1}{N} \sum_{i=1}^N y_i \log(y'_i) + (1 - y_i) \log(1 - y'_i)$$

where φ represents the corresponding model parametrization. Regularization of the model was entailed using: 1) L1 regularization, i.e. linear penalization of the model parametrization 2) dropout, i.e. random drop of nodes at each stage of the training process with a probabilistic rate DR and consecutive weighting of each of the nodes' output with $(1-DR)$ in the prediction inference to yield the expected value of the output.

Explainability techniques for ANNs can be grouped into two categories: gradient-based methods such as saliency [27] and backward propagation methods such as deconvolution [28], guided backpropagation [29], SmoothGrad [30] and layer-wise relevance propagation (LRP) [31]. Saliency is a simple technique that for a given data point identifies the most relevant input features to which the output is most sensitive. The advantage of saliency is the simplicity of the method application. However, it comes with the disadvantage of limited capability to provide explainability, due to its relation to local differential effects only. In comparison, backward propagation methods make use of the graph structure of neural networks by mapping the prediction backwards along each layer using a set of predefined rules and thus can provide better explanations to what made the network arrive at a particular decision [11]. Amongst these methods, LRP provides the advantage of introducing a conservation property during the propagation of relevance values and has shown an excellent benchmark performance [32]. For a specific set of rules, the LRP can be seen as computing a Taylor decomposition of the

relevance at a layer onto its predecessor. This is called deep Taylor decomposition and has been proposed by Samek et al. as the method of choice for the backpropagation rule in LRP [11].

Deep Taylor decomposition is an interpretation of layer-wise relevance propagation when the parameters α and β in the propagation rule are set accordingly [20,31]. These parameters regulate the contribution of positive and negative connections between neurons to the relevance calculation. With $\alpha = 1$ and $\beta = 0$, the relevance projected from a neuron k onto its input neuron j can be written by the following simpler rule which is equivalent to a first order Taylor decomposition:

$$r_{j \leftarrow k} = \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} r_k$$

where a_j is the activation of neuron j and w_{jk}^+ is the positive weight between neurons j and k . Summing $r_{j \leftarrow k}$ over all neurons k to which neuron j contributes to yields the following propagation rule:

$$r_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} r_k$$

All neuron relevance values are propagated layer-wise using this rule from the final output layer until the input, providing the input features with final relevance values.

The overall features importance was calculated as the weighted average of the observations with relation to the confidence of prediction:

$$R(f) = \frac{1}{N} \sum_{i=1}^N \theta_i r_i(f)$$

with $\theta_i = y_i \cdot P(O = 1 | \bar{X}_i) + (1 - y_i)(1 - P(O = 1 | \bar{X}_i))$ where $R(f)$ is the normalized feature rating and $r_i(f)$ is the feature contribution for the given MLP model for observation i using deep Taylor decomposition calculated by the propagation rule presented above.

Models training and validation

The data were randomly split into training- and test sets with a corresponding 4:1 ratio. Mean/mode imputation and feature scaling using zero-mean unit variance normalization based on the training set was performed on both sets. To target class imbalance the training set was randomly sub-sampled to yield uniform class distribution. The models were then tuned using 10-folds cross-validation. The whole process was repeated 50 times (shuffles). Table 2 provides a summary of the tuned hyperparameters for each model.

Performance assessment

The model performance was tested on the test set using receiver-operating-characteristic (ROC)-analysis by measuring the area-under-the-curve (AUC). The performance measure was taken as the median value over the number of shuffles.

Interpretability assessment

The absolute values of the calculated feature importance scores were normalized, i.e. scaled to unit norm, in order to provide comparable feature rating across models: For each sample (each of the 50 shuffles) the calculated importance scores were rescaled to be confined within

Table 2. Summary of hyperparameters tuning.

Model	Hyperparameter	Range
LASSO	C (inverse of regularizer multiplier)	0.10, 0.12, 0.15, 0.18, 0.21, 0.26, 0.31, 0.37, 0.45, 0.54, 0.66, 0.79, 0.95, 1.15, 1.39, 1.68, 2.02, 2.44, 2.95, 3.56, 4.29, 5.18, 6.25, 7.54, 9.10, 10.9, 13.3, 16.0, 19.3, 23.3, 28.1, 33.9, 40.9, 49.4, 59.6, 72.0, 86.9, 105, 126, 153, 184, 222, 268, 324, 391, 471, 569, 687, 829, 1000
Elastic net	L1 ratio	0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95
	Alpha	0.00001, 0.00004, 0.00016, 0.0006, 0.0025, 0.01, 0.04, 0.16, 0.63, 2.5, 10
CatBoost	Tree depth	2, 4
	Learning rate	0.03, 0.1, 0.3
	Bagging temperature	0.6, 0.8, 1.
	L2 leaf regularization	3, 10, 100, 500
	Leaf estimation iterations	1, 2
MLP	Number of hidden neurons	5, 10, 15, 20
	Learning rate	0.001, 0.01
	Batch size	16, 32
	Dropout rate	0.1, 0.2
	L1 regularization ratio	0.0001, 0.001

The table details the hyperparameters and corresponding range that were tuned for each model in the cross-validation process.

<https://doi.org/10.1371/journal.pone.0231166.t002>

the range [0,1] with their sum equal to one. Then, for each feature the mean and standard deviation over the samples (shuffles) were calculated and reported as the final rating measures.

Results

Performance evaluation

All models demonstrated comparable performance for 3 months dichotomized mRS prediction as measured by AUC values on the test set: GLM 0.83, Lasso 0.83, Elastic Nets 0.81, Tree boosting 0.81 and MLP 0.83. While Catboost showed the highest performance, the difference to the other models was very small. For a graphical representation of the models performance on the training and test sets please see [Fig 1](#).

Interpretability analysis

The interpretability analysis demonstrated consistent results across models in terms of the strongest and established predictors: All explainable models rated age and initial NIHSS consistently amongst the most important features. For less important features, results were more varied. The most similar ratings were obtained between the Elastic net and the tree boosting model. The lowest variance amongst feature importance was found for the MLP model. A graphical representation of the results can be found in [Fig 2](#).

Discussion

In the present work, we have used a well-characterized clinical stroke outcome prediction paradigm to compare the ability of modern and traditional machine learning methods to provide explainability of their predictions. In the context of the presented study, both types of ML methods (artificial neural nets and tree boosting) showed comparable performance and similar interpretability patterns for the most important predictors. We corroborated that modern

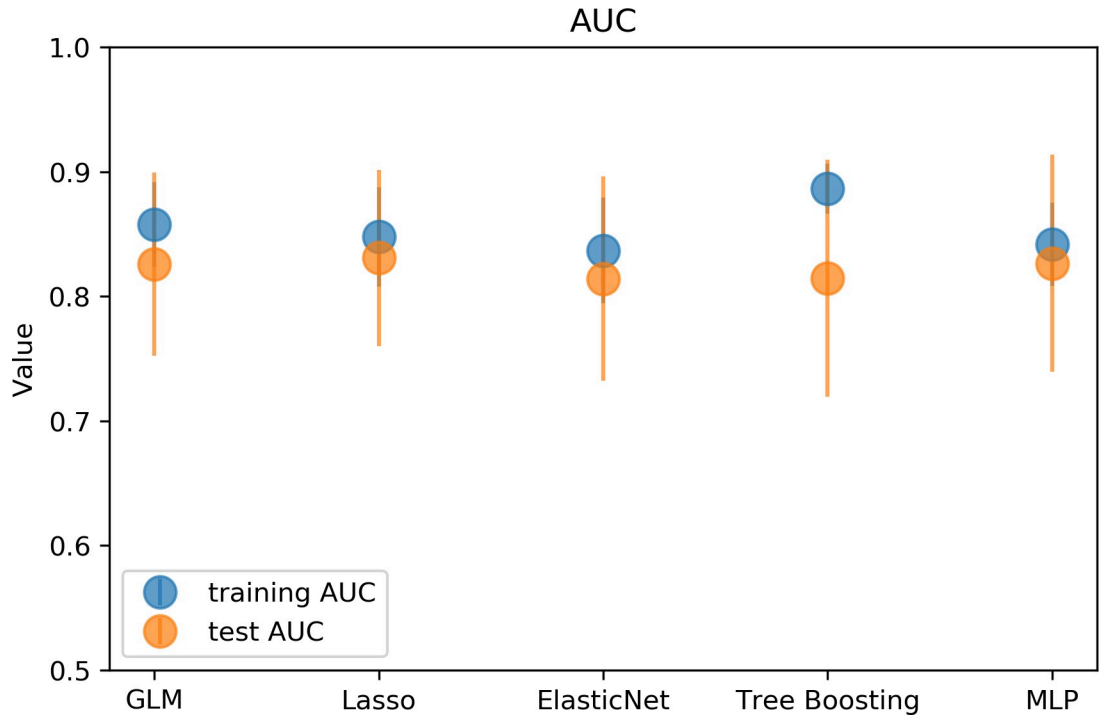


Fig 1. Graphical representation of the model performance results. The graph illustrates the performance of the different models evaluated on the training (blue) and test (orange) sets: generalized linear model (GLM), Lasso, Elastic net, Tree Boosting and multilayer perceptron (MLP). The markers show the median AUC over 50 shuffles and the error bars represent interquartile range (IQR). All models showed a similar median AUC around 0.82. The largest difference in performance between training and test set, indicating potential overfitting, was observed for the Catboost model.

<https://doi.org/10.1371/journal.pone.0231166.g001>

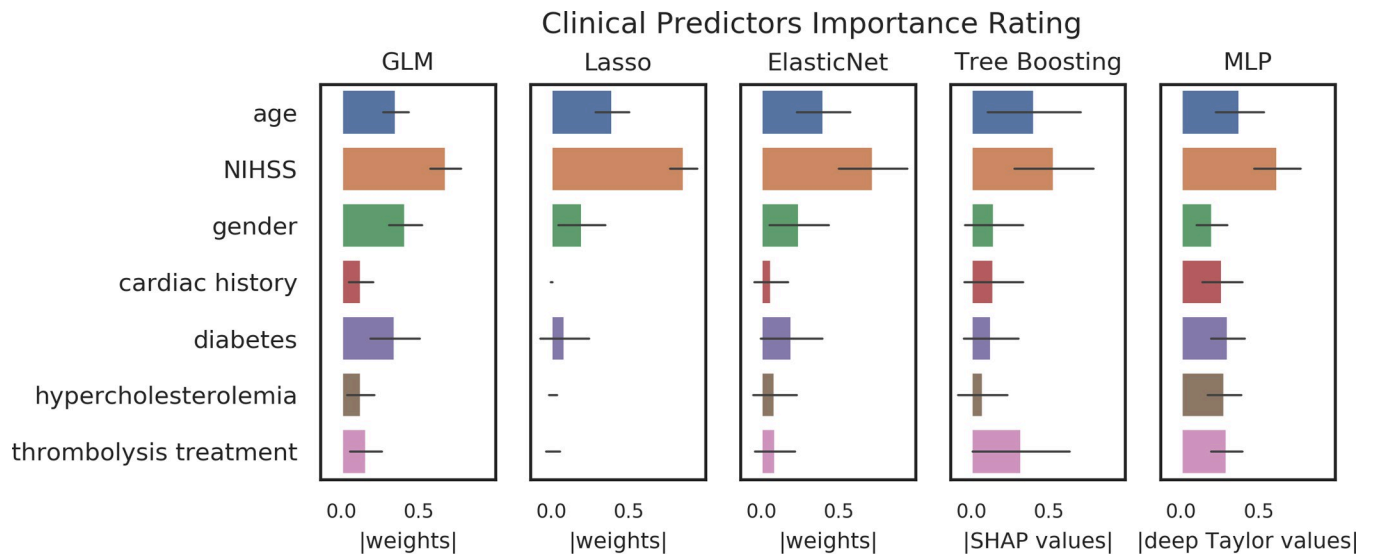


Fig 2. Graphical representation of the feature importance. The figure illustrates the features rating derived from the model-tailored interpretability methods for generalized linear model (GLM), Lasso, Elastic net, Catboost and multilayer perceptron (MLP). All models rated age and initial NIHSS consistently amongst the most important features. For less important features, results were more varied. For logistic regression techniques the results are given in weights, for Catboost in Shap(ley) values and for MLP in deep Taylor values that were normalized to the range [0,1]. The bar heights represent means and error bars represent standard deviation over samples (shuffles).

<https://doi.org/10.1371/journal.pone.0231166.g002>

techniques are not necessarily black boxes, but are able to provide a reliable assessment of feature importance comparable to their traditional counterparts for clinical prediction models.

In contrast to other domains, models in healthcare require higher levels of safety given that patients' life and health is at stake [33]. Here, the explainability of the predictions is a highly important criterion to enable it. Unfortunately, explainability in the modeling context is an ill-defined term that can also have other meanings and several other terms such as interpretability and transparency are in use [34]. A comprehensive overview is beyond the scope of the current work, but we would like to introduce two examples. Doran et al. define interpretability as methodological explainability, e.g. the weights of a linear regression algorithm, in contrast to comprehensibility which is a symbolic representation of an output [35]. This view focuses on different users. Interpretability methods can aid developers in the development process, e.g. as a means to find and avoid mistakes. Comprehensible explainability on the other hand refers to how the results are presented to the user in the final product. In healthcare, the users are healthcare professionals with very limited understanding of the technical background of prediction models. Thus, the exact nature of this presentation must be determined—on a case by case basis—for each product. In some cases, more technical presentations as also shown in Fig 2 might suffice. For others, it might be necessary to translate the rankings into easier to understand formats, e.g. categories (“very important” vs. “important” vs “unimportant”). To determine these characteristics is the domain of User experience/User interface (Ux/UI) analysis, where a thorough testing with users must be performed. This view defines interpretability as a sub-category of explainability. This view defines interpretability as a sub-category of explainability. Others see a distinction. Rudin defines interpretability as an attribute of a method, i.e. a method which inherently provides information about feature importance, such as the weights of linear regression [36]. Explainability on the other hand describes a model which is used to approximate the original model to derive a surrogate interpretability. Such methods can be tailored to one specific original black box algorithm, or can be generalized like the LIME algorithm [37]. We would like to stress that no standardization of these terms currently exists. Thus, in the presented work, explainability is mainly examined from a clinical point-of-view, highlighting the ability of humans to understand which clinical features drive the prediction. This is important, as a major goal of clinical predictive modeling is the development of clinical decision support systems (CDSS) aiding healthcare professionals in their clinical decision making, predicting diagnoses, risks, and outcomes [2,3]. Here, it is important to keep in mind that the requirements for CDSSs go far beyond the model performance [33]. It is established that CDSSs for the clinical setting need to exhibit proven safety [13]. A crucial part of the safety assessment of ML/AI products is to understand why they do what they do, but, more importantly, to understand why and when they might *not* do what is intended. This is important in the light of the increasing awareness of potential biases in models used for healthcare discriminating based on for example sex and gender or ethnicity [38]. Another reason is automation bias—an established cognitive bias—where users tend to believe what a machine is outputting without reflecting on the output [2]. Providing model explainability might mitigate this bias. Thus, it is very likely that future regulatory requirements, e.g. by European MDR and US FDA, will include requests for explainability [39]. Here, our results are highly encouraging. Modern ML methods that are able to provide the potentially highest performance can be combined with methods of explainability and the results are comparable to the established methods for traditional techniques. Thus, researchers and developers are no longer faced with the potential trade-off between lower performance vs. explainability.

However, not only regulatory bodies will require explainability. From the physician point-of-view, black-box approaches might be unacceptable [13,33]. Clinical guidelines for CDSS may therefore profit from explainable predictions. While it has been argued that we have

accepted similar uncertainty in medical decision making to date and accuracy alone can be sufficient [40], we would argue that explainability is a must-have when it can be added without limiting the accuracy, as our results suggest. Nonetheless, explainability is a supportive tool and is not a substitute for rigorous clinical validation of any CDSS[40].

We have focused in our work on two promising techniques, namely artificial neural nets and tree boosting. ANNs have shown highly promising results in several areas of healthcare such as medical imaging, information extraction from medical texts and electronic health records, and combining several types of input into one predictive model [5]. Also tree boosting has shown high performance across several medical domains [41]. Tree boosting algorithms are also much easier to train than artificial neural nets and their performance is quite immune to feature scaling and collinearity issues. Another major advantage of tree boosting in healthcare is scalability [42] and thus it is also suited for big data analytics, for example data mining from electronic health records (EHR). Here, tree boosting can achieve comparable performance to deep learning techniques [43]. As evidenced by the above, tree boosting and ANNs represent very versatile and well performing modern ML algorithms in healthcare. Thus, our work is of high practicality for future research and for clinical decision support development.

The main focus of our work was the comparison of explainability in a well-characterized prediction paradigm and not a comparison of performance. It is not surprising that both the traditional and the modern ML methods achieved comparable performance in our dataset. Given the simplicity of the classification problem and the limited dataset, traditional methods are sufficient to capture the relationship of the features to the prediction and complex methods may easily result in overfitting. It is, however, important to note that interpretability without a certain performance level is meaningless: A randomly classifying classifier cannot provide reliable feature importance. If, however, the performance of modern ML methods were considerably higher and the methods' explainability were to be more reliable, it cannot be determined whether this increase resulted due to a better explainability method or due to a performance increase. Thus, the simplicity of the paradigm we chose is well suited to compare explainability, as the performance is comparable and feature ratings provide a straight-forward result that can be assessed against domain knowledge. Had the performance varied considerably, interpretation of the rankings might have been severely impaired. With regard to our explainability analysis, several more observations are noteworthy. As there is no gold-standard to interpret rankings it can only be performed against domain-knowledge and through replication studies. While we know from previous studies that age, NIHSS and thrombolysis are important predictors to predict stroke outcome (with age and NIHSS being the two strongest) [16–19], it is crucial to include the specifics of the dataset into the interpretation. The median NIHSS of the sample was only 3 and only around 31% of patients received thrombolysis, meaning that many of the patients had smaller—less serious—stroke events. As a consequence, the potential effect of thrombolysis is limited in our sample. Thus we would—like in the above mentioned previous works—expect that age and NIHSS drive the prediction. And indeed, all rankings gave these two very high importance, with the exception of the GLM ranking they were the two most important predictors. The ranking of the lesser predictors, however, varied relatively strongly. Interestingly, elastic net provided the ranking which is most similar to the one provided by tree boosting. From a domain perspective, the most reliable and complete ranking was provided by the tree boosting model, ranking age and NIHSS unequivocally on top, with thrombolysis being slightly more important than the other features. While the MLP gave age and NIHSS the expected high importance, it ranked the presence of diabetes similarly strong. A similar ranking for diabetes can also be observed in the logistic regression models. Although diabetes is known to be an important predictor for bad stroke outcome [44], a feature importance score that is at a similar level as age is unexpected. Another striking difference is the high

relative importance given to sex by the logistic regression models, which is absent in the rankings provided by the modern methods. Taken together, we observed promising consistent findings, where all methods corroborated the importance of age and NIHSS for stroke outcome prediction. At the same time, we saw distinct differences for diabetes and sex which cannot be explained sufficiently at the current time point. In light of these findings, we certainly do not claim that the explanations provided by the modern methods should be taken without further validation. Our work established that rankings can be obtained for modern machine learning methods and that these rankings are compatible with clinical interpretation, especially regarding the main predictors. The differences between the rankings, however, must be the subject of further research. Here, it must be mentioned that for ANNs multiple other methods than Taylor decomposition exist, which should also be further tested in the future—a task which was beyond the scope of the current work.

Given the aforementioned trade-off between performance and explainability, a distinction between traditional and modern techniques seems justifiable. It carries with it, however, the risk that modern methods are overhyped and used where traditional techniques might perform best. As our results suggest that also modern techniques provide explainability, we would argue that this distinction is irrelevant. Once all important methods for clinical predictive modeling provide validated feature importance we should simply choose the method which seems best suited for the prediction task at hand. We believe that this will greatly facilitate the development of clinical decision support systems.

Our work has several limitations. First, we used only one dataset. Here, our results are promising, but clearly more analyses are warranted to compare rankings provided by modern ML methods with rankings provided by traditional ML methods. Second, to allow comparison with traditional methods, we used a paradigm that utilizes only clinical values. We encourage future works evaluating explainability provided for other data modalities such as imaging.

Conclusions

For the first time, we established in an empirical analysis on clinical data that modern machine learning methods can provide explainability which is compatible with domain knowledge interpretation and traditional method rankings. This is highly encouraging for the development of explainable clinical predictive models. Future work should validate the explainability methods, further explore the differences between them, and test different predictive modeling frameworks including multiple modalities.

Author Contributions

Conceptualization: Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey.

Data curation: Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach.

Formal analysis: Esra Zihni, Vince Istvan Madai, Michelle Livne.

Funding acquisition: Jochen B. Fiebach, Dietmar Frey.

Investigation: Esra Zihni, Vince Istvan Madai, Michelle Livne, Dietmar Frey.

Methodology: Esra Zihni, Vince Istvan Madai, Michelle Livne, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey.

Project administration: Vince Istvan Madai, Jochen B. Fiebach, Dietmar Frey.

Resources: Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey.

Supervision: Vince Istvan Madai, Michelle Livne, Dietmar Frey.

Validation: Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil.

Visualization: Vince Istvan Madai, Michelle Livne.

Writing – original draft: Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey.

Writing – review & editing: Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey.

References

1. Khamparia A, Singh KM. A systematic review on deep learning architectures and applications. *Expert Syst.* 2019; 36: e12400. <https://doi.org/10.1111/exsy.12400>
2. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* 2019; 28: 231–237. <https://doi.org/10.1136/bmjqs-2018-008370> PMID: 30636200
3. Ashrafian H, Darzi A. Transforming health policy through machine learning. *PLOS Med.* 2018; 15: e1002692. <https://doi.org/10.1371/journal.pmed.1002692> PMID: 30422977
4. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.* 2018; 19: 1236–1246. <https://doi.org/10.1093/bib/bbx044> PMID: 28481991
5. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med.* 2019; 25: 24–29. <https://doi.org/10.1038/s41591-018-0316-z> PMID: 30617335
6. Luo L, Li J, Liu C, Shen W. Using machine-learning methods to support health-care professionals in making admission decisions. *Int J Health Plann Manage.* 2019; 34: e1236–e1246. <https://doi.org/10.1002/hpm.2769> PMID: 30957270
7. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang S-W, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLOS ONE.* 2019; 14: e0221202. <https://doi.org/10.1371/journal.pone.0221202> PMID: 31442238
8. Livne M, Boldsen JK, Mikkelsen IK, Fiebach JB, Sobesky J, Mouridsen K. Boosted Tree Model Reforms Multimodal Magnetic Resonance Imaging Infarct Prediction in Acute Stroke. *Stroke.* 2018; 49: 912–918. <https://doi.org/10.1161/STROKEAHA.117.019440> PMID: 29540608
9. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access.* 2018; 6: 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
10. Nathans LL, Oswald FL, Nimon K. Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. 2012; 17: 19.
11. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. *Digit Signal Process.* 2018; 73: 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
12. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.* New York, NY, USA: ACM; 2018. pp. 559–560. <https://doi.org/10.1145/3233547.3233667>
13. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018; 320: 2199–2200. <https://doi.org/10.1001/jama.2018.17163> PMID: 30398550
14. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLOS Med.* 2018; 15: e1002689. <https://doi.org/10.1371/journal.pmed.1002689> PMID: 30399149
15. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems 30.* Curran Associates, Inc.; 2017. pp. 4765–4774. Available: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
16. Khosla A, Cao Y, Lin CC-Y, Chiu H-K, Hu J, Lee H. An integrated machine learning approach to stroke prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM; 2010. pp. 183–192.

17. 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. <https://doi.org/10.1371/journal.pone.0088225> PMID: 24520356
18. Weimar C, Roth MP, Zillesen G, Glahn J, Wimmer MLJ, Busse O, et al. Complications following Acute Ischemic Stroke. *Eur Neurol*. 2002; 48: 133–140. <https://doi.org/10.1159/000065512> PMID: 12373029
19. Parsons MW, Christensen S, McElduff P, Levi CR, Butcher KS, De Silva DA, et al. Pretreatment diffusion- and perfusion-MR lesion volumes have a crucial influence on clinical response to stroke thrombolysis. *J Cereb Blood Flow Metab Off J Int Soc Cereb Blood Flow Metab*. 2010 [cited 16 Feb 2010]. <https://doi.org/10.1038/jcbfm.2010.3> PMID: 20087363
20. Montavon G, Lapuschkin S, Binder A, Samek W, Müller K-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit*. 2017; 65: 211–222. <https://doi.org/10.1016/j.patcog.2016.11.008>
21. Hotter B, Pittl S, Ebinger M, Oepen G, Jegzentis K, Kudo K, et al. Prospective study on the mismatch concept in acute stroke patients within the first 24 h after symptom onset - 1000Plus study. *BMC Neurol*. 2009; 9: 60. <https://doi.org/10.1186/1471-2377-9-60> PMID: 19995432
22. O'Brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual Quant*. 2007; 41: 673–690. <https://doi.org/10.1007/s11135-006-9018-6>
23. Nimon KF, Oswald FL. Understanding the Results of Multiple Linear Regression: Beyond Standardized Regression Coefficients. *Organ Res Methods*. 2013; 16: 650–674. <https://doi.org/10.1177/1094428113493929>
24. Miles J. Tolerance and Variance Inflation Factor. *Wiley StatsRef: Statistics Reference Online*. American Cancer Society; 2014. <https://doi.org/10.1002/9781118445112.stat06593>
25. catboost: Catboost Python Package. Available: <https://catboost.ai>
26. Csáji BC. Approximation with Artificial Neural Networks. 2001; 45.
27. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv13126034 Cs*. 2013 [cited 2 Sep 2019]. Available: <http://arxiv.org/abs/1312.6034>
28. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer Vision—ECCV 2014*. Cham: Springer International Publishing; 2014. pp. 818–833. https://doi.org/10.1007/978-3-319-10590-1_53
29. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net. *ArXiv14126806 Cs*. 2015 [cited 23 Feb 2020]. Available: <http://arxiv.org/abs/1412.6806>
30. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. *ArXiv170603825 Cs Stat*. 2017 [cited 2 Sep 2019]. Available: <http://arxiv.org/abs/1706.03825>
31. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*. 2015; 10: e0130140. <https://doi.org/10.1371/journal.pone.0130140> PMID: 26161953
32. Samek W, Binder A, Montavon G, Lapuschkin S, Müller K-R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans Neural Netw Learn Syst*. 2017; 28: 2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820> PMID: 27576267
33. Yu K-H, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf*. 2019; 28: 238–241. <https://doi.org/10.1136/bmjqs-2018-008551> PMID: 30291179
34. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable Machine Learning for Scientific Insights and Discoveries. *ArXiv190508883 Cs Stat*. 2019 [cited 3 Sep 2019]. Available: <http://arxiv.org/abs/1905.08883>
35. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *ArXiv171000794 Cs*. 2017 [cited 3 Sep 2019]. Available: <http://arxiv.org/abs/1710.00794>
36. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019; 1: 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
37. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16*. San Francisco, California, USA: ACM Press; 2016. pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>
38. Nelson GS. Bias in Artificial Intelligence. *N C Med J*. 2019; 80: 220–222. <https://doi.org/10.18043/ncm.80.4.220> PMID: 31278182
39. johner-institut/ai-guideline. In: GitHub [Internet]. [cited 4 Oct 2019]. Available: <https://github.com/johner-institut/ai-guideline>

40. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep.* 2019; 49: 15–21. <https://doi.org/10.1002/hast.973> PMID: 30790315
41. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O, AME Big-Data Clinical Trial Collaborative Group W on BO. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med.* 2019;7. <https://doi.org/10.21037/atm.2018.12.26>
42. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *ArXiv160302754 Cs.* 2016; 785–794. <https://doi.org/10.1145/2939672.2939785>
43. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, et al. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci Rep.* 2019; 9: 1–10. <https://doi.org/10.1038/s41598-018-37186-2>
44. Lau L-H, Lew J, Borschmann K, Thijs V, Ekinci EI. Prevalence of diabetes and its effects on stroke outcomes: A meta-analysis and literature review. *J Diabetes Investig.* 2019; 10: 780–792. <https://doi.org/10.1111/jdi.12932> PMID: 30220102

2.2 A precision medicine framework for personalized simulation of hemodynamics in cerebrovascular disease³³

Die in der vorangegangenen Arbeit vorgelegten Ergebnisse bahnten den Weg, um ein Framework zur Simulation der Hämodynamik der Hirngefäße zu entwickeln und hiermit erstmals für zerebrovaskulären Erkrankungen einen Simulations-basierten und personalisierten Ansatz zu entwickeln.

Wie bereits beschrieben, stellen zerebrovaskuläre Erkrankungen, insbesondere der Schlaganfall, eine große Herausforderung für das Gesundheitssystem dar. Ein wichtiger Biomarker für die Diagnostik und Therapie dieser Erkrankungen ist die Durchblutung des Hirngewebes, die zerebrale Hämodynamik. Zur Messung und Quantifizierung der zerebralen Hämodynamik stehen momentan ausschließlich invasive und potenziell gesundheitsschädliche oder die Behandlungszeit verlängernde Methoden zur Verfügung.

In der folgenden Arbeit stellen wir eine simulationsbasierte Methode vor, die eine non-invasive Berechnung der zerebralen Hämodynamik ermöglicht. Auf der Grundlage der patientenindividuellen Gefäßkonfiguration der strukturellen Gefäßdarstellung haben wir ein System implementiert, welches als erste Schritte die Segmentierung und Annotation von Hirngefäßen aus der strukturellen Bildgebung ermöglicht. Für die Annotation wurde eine grafische 3D-Benutzeroberfläche implementiert.

Der resultierende annotierte Gefäßbaum wird dann in eine 0-dimensionale Simulationsmodellierung der zerebralen Hämodynamik überführt, wobei hierfür eine modifizierte Knotenanalyse verwendet wurde. Die resultierende Simulation ermöglicht die Identifizierung von schlaganfallgefährdeten, sogenannten vulnerablen Bereichen. Daneben können Veränderungen aufgrund unterschiedlicher systemischer Blutdrücke simuliert werden, um deren Einfluss auf die lokale Hirndurchblutung zu evaluieren.

Darüber hinaus wurde eine Sensitivitätsanalyse implementiert, die die Live-Simulation von Veränderungen ermöglicht, um Verfahren und Krankheitsverläufe zu simulieren. In der durchgeführten explorativen Analyse von 67 Patienten konnte gezeigt werden, dass die Simulation für die Detektion von Perfusionsveränderungen in der klassischen Perfusionsbildgebung eine hohe Spezifität und eine geringe bis mittlere Sensitivität aufweist.


Damit hat der hier vorgestellte Ansatz der Präzisionsmedizin mittels Verwendung eines neu entwickelten Biomarkers das Potenzial, die Anwendung gesundheitsschädlicher und zeit- und ressourcenintensiver Perfusionsmethoden mit einer individuellen Simulation der Hirndurchblutung zu ersetzen. Für die breite Anwendung der hier vorgestellten Simulation sind weitere Validierungsschritte erforderlich.

RESEARCH

Open Access



A precision medicine framework for personalized simulation of hemodynamics in cerebrovascular disease

Dietmar Frey^{1,4*} , Michelle Livne¹, Heiko Leppin¹, Ela M. Akay^{1,4}, Orhun U. Aydin^{1,4}, Jonas Behland^{1,4}, Jan Sobesky^{2,3}, Peter Vajkoczy⁴ and Vince I. Madai^{1,5}

*Correspondence:

dietmar.frey@charite.de

¹ Charite Lab for Artificial Intelligence in Medicine, Department of Neurosurgery, Charité University Medicine Berlin, Chariteplatz 1, 10115 Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Background: Cerebrovascular disease, in particular stroke, is a major public health challenge. An important biomarker is cerebral hemodynamics. To measure and quantify cerebral hemodynamics, however, only invasive, potentially harmful or time-to-treatment prolonging methods are available.

Results: We present a simulation-based approach which allows calculation of cerebral hemodynamics based on the patient-individual vessel configuration derived from structural vessel imaging. For this, we implemented a framework allowing segmentation and annotation of brain vessels from structural imaging followed by 0-dimensional lumped simulation modeling of cerebral hemodynamics. For annotation, a 3D-graphical user interface was implemented. For 0D-simulation, we used a modified nodal analysis, which was adapted for easy implementation by code. The simulation enables identification of areas vulnerable to stroke and simulation of changes due to different systemic blood pressures. Moreover, sensitivity analysis was implemented allowing the live simulation of changes to simulate procedures and disease progression. Beyond presentation of the framework, we demonstrated in an exploratory analysis in 67 patients that the simulation has a high specificity and low-to-moderate sensitivity to detect perfusion changes in classic perfusion imaging.

Conclusions: The presented precision medicine approach using novel biomarkers has the potential to make the application of harmful and complex perfusion methods obsolete.

Keywords: Simulation, Precision medicine, MR imaging, Hemodynamics, Cerebrovascular disease, Medical software, Segmentation, Annotation, Machine learning

Background

Cerebrovascular disease, and in particular stroke, is a major public health challenge. It is a leading cause of death and disability [1]. While there have been advances in prevention and treatment in the past—e.g., mechanical thrombectomy for acute stroke treatment—the overall prevention and treatment results still remain poor [2]. A potential



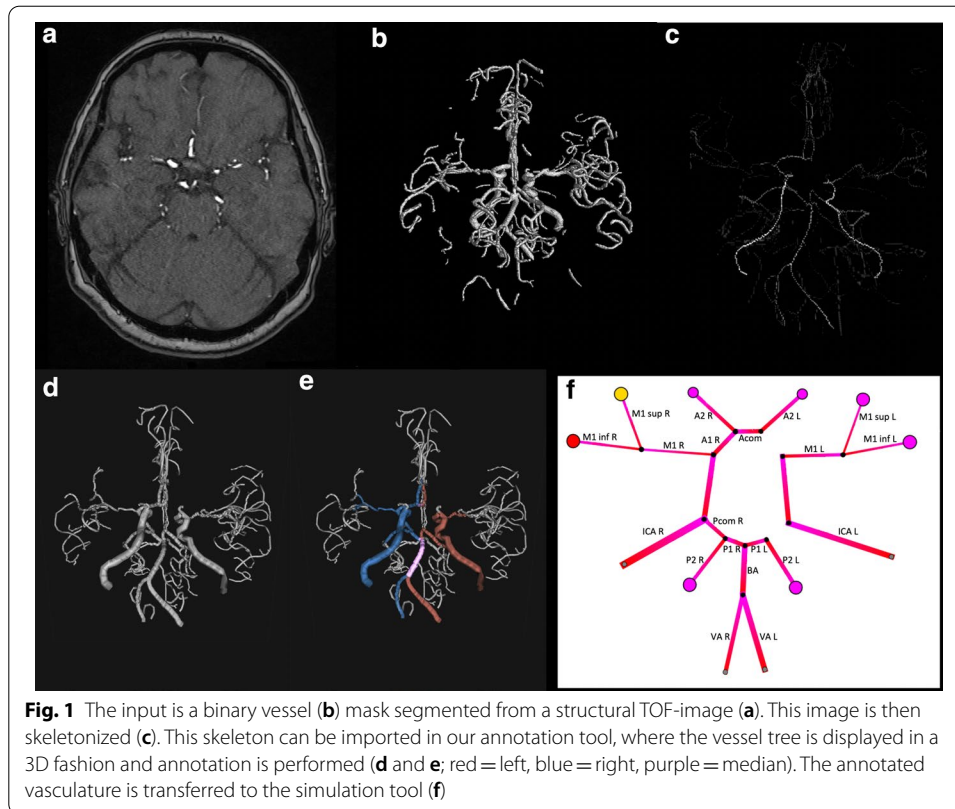
© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

game-changer of stroke treatment success is precision medicine [3, 4]. It aims to provide personalized therapy recommendations based on the individual features of the patient. It utilizes today's plethora of available patient data as well as mathematical modeling to offer individualized predictions for patients [4]. While highly promising, precision medicine relies on the presence of informative data allowing the differentiation of pathology patterns [4, 5]. In cerebrovascular disease, important information about the severity of stroke risk and potential response to treatment is encoded in individual pathophysiological parameters—in biomarkers—which can be recorded to aid decision-making. Here, one of the most important parameters is the hemodynamic status [6]. This biomarker is already used in a precision medicine approach to identify individual patients benefiting from thrombolysis beyond the currently established treatment time windows which is crucial since often treatment is denied due to time constraints [7]. In chronic cerebrovascular disease, it might aid by identifying areas which are highly vulnerable to stroke [8, 9]. In the clinical setting, however, this data is only available using specialized methodologies, i.e., Dynamic Susceptibility-weighted Contrast-enhanced Magnetic Resonance Imaging (DSC-MRI) perfusion, computed-tomography (CT)-perfusion, arterial spin labeling (ASL) perfusion or functional MRI [10–14]. These techniques may harm patients through contrast agents, significantly prolong the time to treatment and lead to increased costs. Also, standardization of these complex methods is highly challenging [6, 15, 16].

An alternative approach to derive biomarkers for precision medicine is the transformation of routinely acquired data by mechanistic simulations [17]. These simulations integrate domain knowledge by mathematically describing known disease-driving core processes [17]. Interestingly for cerebrovascular disease, several works in the past have developed general mechanistic simulations of the blood flow in the brain [18, 19]. These simulations have the potential to become a contrast agent-free biomarker of hemodynamics for the diagnosis and treatment of cerebrovascular diseases. However, for these simulations, personalization on an individual patient level is still pending making it not applicable in a clinical setting.

Thus, the novel idea presented in this work is a software framework to transform routine structural vessel imaging data as an input to a mechanistic simulation of individual hemodynamics for a given patient. The unique vessel configuration of each patient can be used to simulate hemodynamics to potentially identify areas that are vulnerable in case of stenosis and occlusion. Several use cases can be envisioned for such a framework. It could allow assessment of stroke risk, pre-operative simulation of interventional success like thrombectomy in acute stroke, preventive or therapeutic endarterectomy and stenting of brain-supplying vessels, respectively. Another highly interesting, if rather rare case is the simulation of the outcome of extracranial–intracranial (EC–IC) bypass surgery, e.g., in Moya-Moya disease. Here, there is a special need to predict the success of the surgery [20]. Lastly, the simulation information could be used for the prediction of stroke outcome in conjunction with other clinical and imaging parameters enabling clinicians with an objective criterion for decision support in the acute setting.

Thus, the objective of the presented work was to provide a framework allowing the incorporation of individual structural vessel data to simulate areas of higher hemodynamic vulnerability as a disease biomarker. For this purpose, we developed a pipeline



consisting of the following sequential steps: (1) segmentation of vessel information from structural data, in our case from time-of-flight (TOF) magnetic resonance imaging (MRI). (2) Annotation of the vessel tree with an easy-to-use graphical user interface (GUI). And (3) simulation where results can be inspected, and different blood pressure scenarios can be simulated by the user.

The simulation was implemented as a steady-state zero-dimensional lumped model of the Circle of Willis (CoW) and major brain artery circulation. We included individual vessel resistances by 1-dimensional calculation using the individual length and the width of the arteries from patient structural vessel imaging. Hemodynamic measures were calculated using an adapted version of the modified nodal analysis (MNA) [21] coined AMNA, where we simplify the solution of the matrix equations facilitating easier implementation and faster runtime. We show the implemented framework in detail—including a video of the annotation process—and perform an exploratory visual analysis to compare simulation results with perfusion imaging in 67 patients with cerebrovascular disease.

Results

The framework architecture

We developed a novel framework for enabling processing of routine imaging (DICOM) into a simulation tool that provides additional information about the state of the hemodynamics of an individual patient.

Figure 1 gives an overview of the subsequent steps of the framework:

1. A structural DICOM TOF-image (Fig. 1a) is processed via segmentation into a binary vessel mask (Fig. 1b).
2. This image is then skeletonized (Fig. 1c).
3. The vessel skeleton is imported in our annotation tool, in which the vessel tree is displayed in a 3-dimensional fashion and
4. Vessel annotation is performed (Fig. 1d and e).
5. The annotated vasculature is then transferred to the simulation tool (Fig. 1f).

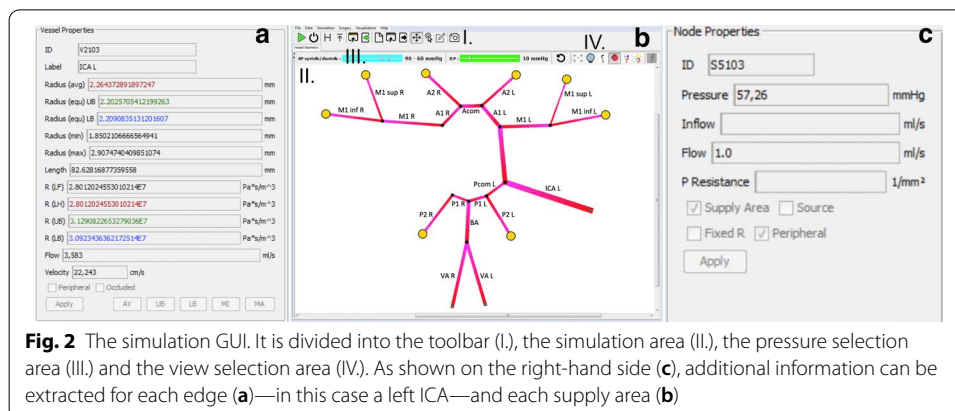
The annotation module

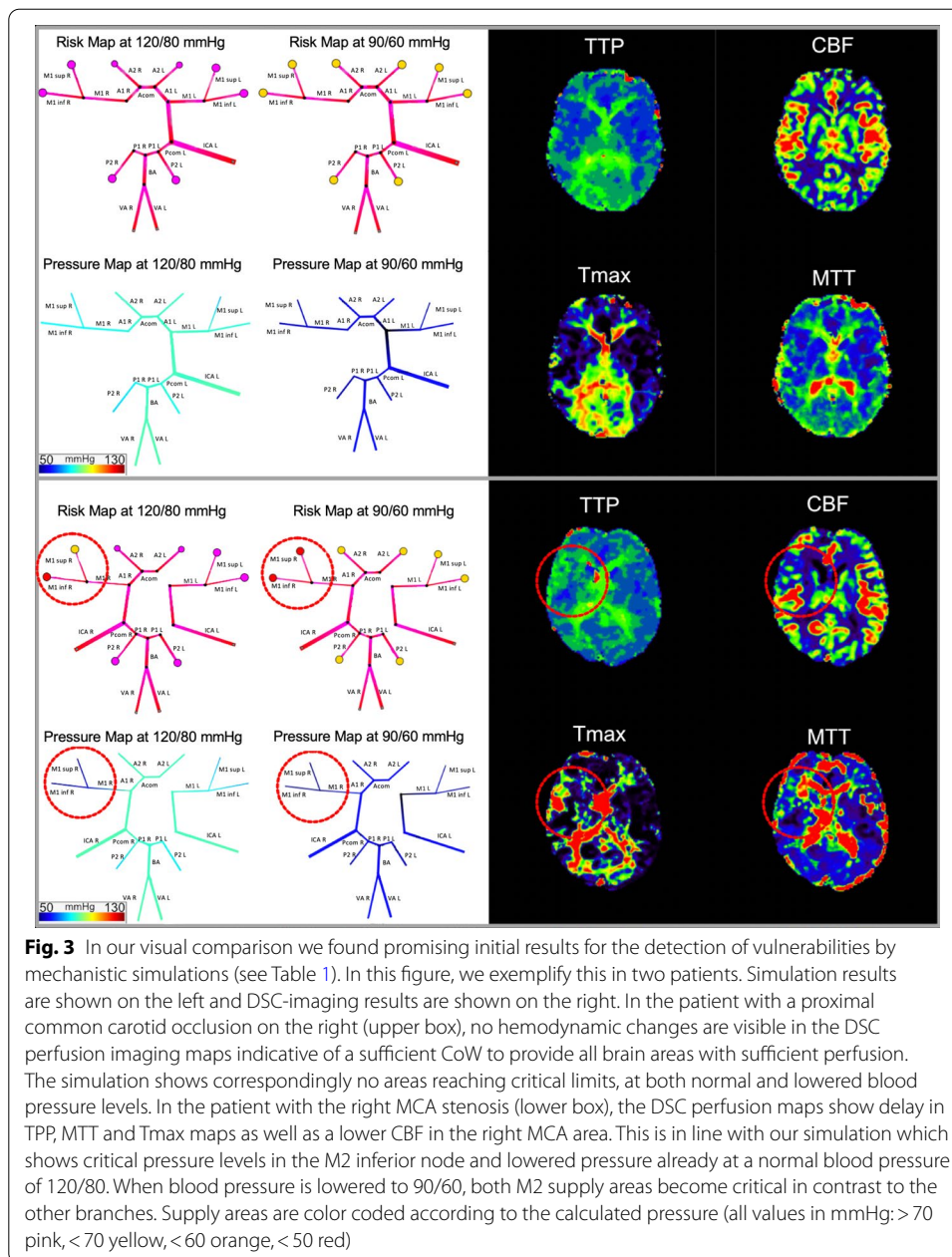
The annotation tool is made up of two main components. The segment annotation area, where the vessel segments can be chosen and the 3D view, where the imported vessel tree can be manipulated. We implemented 22 segments of brain-supplying arteries, namely 3 for the carotid artery (common, internal and CoW segment), 7 for the MCA (M1, M2 superior, M2 inferior, M3 superior superior, M3 superior inferior, M3 inferior superior, M3 inferior inferior), 5 for the ACA (A1, A2, A3 inferior and A3 superior), 4 for the PCA (P1, P2 and P3 inferior and P3 superior), as well as the basilar artery, the vertebral artery and the anterior and posterior communicating arteries. There is the option to add bypass vessels or collateral vessels manually, e.g., for planning of interventions or surgical procedures.

Next to segment also the following additional labels can be chosen: “pre-occlusion”, “post-occlusion”, and “occlusion” (in case pre- or post-occlusion cannot be determined with certainty). To record occlusion next to the type of vessel is important for the following simulation step, as subsequently the simulation will ignore segments with this flag. A video footage of the annotation process was uploaded to zenodo [22].

The simulation module

The simulation itself consists of a graphical user interface, that is divided functionally in the toolbar (Fig. 2b I.), the simulation area (Fig. 2b II.), the pressure selection area (4B III.) and the view selection area (Fig. 2b IV.). In the toolbar, simulations can be loaded as well as the type of resistance calculation (here, the default is the resistance calculation presented in the methods section). In 4B III., the pressure boundary conditions can be chosen, for one the blood pressure which determines the driving pressure of the whole system.





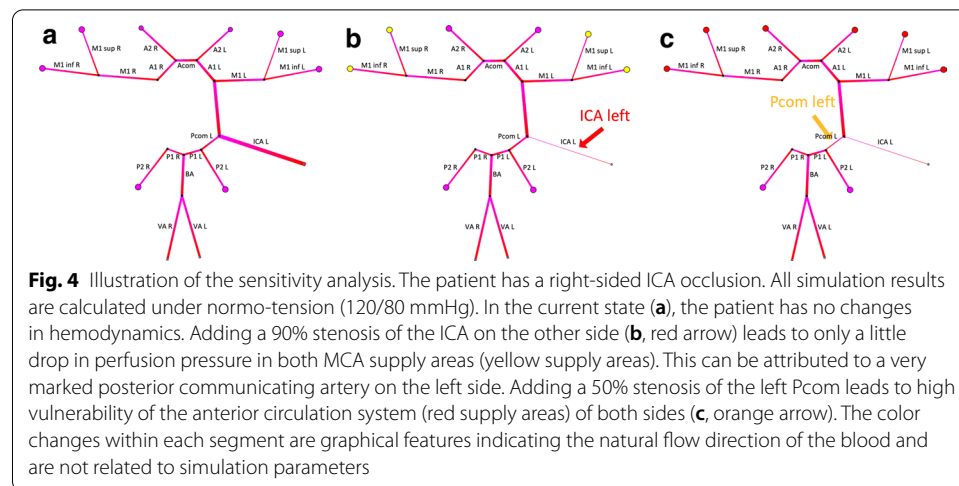
In addition, also the intracranial pressure, which is kept constant for our use case at physiological parameters, but can be increased to simulate conditions with increased intracranial pressure, e.g., in hemorrhagic stroke. In the view selector area, in 4B IV., the normal view and the pressure view can be chosen (please see examples in Fig. 3). Finally, in the simulation area, the individualized simulation of the vasculature can be inspected. Via right click it is possible to derive more information about an edge or a node (Fig. 2a and c, respectively). Supply areas are color coded according to the calculated pressure (all values in mmHg: > 70 pink, < 70 yellow, < 60 orange, < 50 red). Below 50 mmHg we consider an area being vulnerable to ischemia due to the limits of cerebral auto-regulation (see Eq. 5).

Table 1 Detailed results of the visual vulnerability analysis

	ACA left	ACA right	ACA averaged	MCA left	MCA right	MCA averaged	PCA left	PCA right	PCA averaged
Sensitivity	0.50	0.33	0.42	0.49	0.46	0.47	0.07	0.25	0.16
Specificity	0.92	0.97	0.94	1	0.95	0.98	0.92	0.97	0.94

Results are given in rows 1 and 2 for each rated region, including averaged values (in bold) for each perfusion territory (ACA, MCA, PCA). The visual analysis revealed a very high specificity and a moderate-to-low sensitivity for the simulation to detect vulnerability defined by visual DSC-rating. The highest sensitivity was found for the MCA region, followed by the ACA and PCA regions

ACA anterior cerebral artery, MCA middle cerebral artery, PCA posterior cerebral artery



Comparison of DSC perfusion imaging and simulation results

67 patients (mean age 57 years) were included. 68% had previous stroke and 18% had a previous transitory ischemic attack (TIA). The visual analysis revealed a very high specificity and a moderate-to-low sensitivity for the simulation to detect vulnerability defined by visual DSC-rating. Results averaged over both hemispheres for each perfusion territory were (sensitivity/specificity): ACA 0.42/0.94, MCA 0.47/0.98, PCA 0.16/0.94. For detailed results please see Table 1.

Sensitivity analysis

An example of the sensitivity analysis is shown in Fig. 4, where we give an example how the still normal hemodynamics in a patient with a missing ICA would be affected by additional vessel stenosis. The sensitivity analysis successfully allows individualized simulations of potential disease progressions and blood pressure scenarios.

Discussion

We present the first comprehensive precision medicine pipeline for cerebrovascular disease that is capable to process routine stroke imaging DICOM images for simulation of various boundary conditions to identify brain areas vulnerable for ischemia.

Our framework provides annotation of the arterial vasculature derived from neuroimaging followed by zero-dimensional individualized simulation of brain hemodynamics. Implementation was performed by decreasing the computational burden by a modified MNA, the development of an easy-to-use web user interface-frontend for annotation and a java-based cross-platform simulation tool. An exploratory validation analysis comparing our simulation results with DSC perfusion in patients with steno-occlusive disease revealed promising initial results for the simulation-based detection of vulnerable areas. Our results suggest that mechanistic simulation of blood flow derived from routine structural imaging can serve as an individual biomarker for patients with cerebrovascular disease and might be an alternative to complex and potentially harmful perfusion techniques.

Stroke is a complex disease with a dynamic progression. The initial infarct area—characterized by rapid neuronal loss—is called the core, which is surrounded by tissue that is slowly surrendering to ischemia but is still salvageable. The latter area is coined the penumbra and defines the therapeutic target in acute stroke management [6]. For the understanding of stroke and its implications on treatment strategy, it is essential that the speed of the penumbra-to-core transformation varies greatly and is highly individual. In particular, for weighing benefit and risk for stroke treatment the high interindividual variance of brain cell death is crucial: some patients do not have salvageable brain tissue already a few hours after stroke, whereas in others penumbral tissue was found up to 17 h after stroke [23]. This highly individual and variant stroke progression stands in stark contrast to the current “one-size-fits-all” treatment approach in stroke where patients receive treatment based on guidelines usually only within a predefined time window of up to 4.5 h after stroke [6, 24]. These time windows were established by a statistical benefit-to-risk calculation after lumping together all stroke patients with assumed common pathophysiology.

While it is true that there is a net profit for patients when treatment is applied within this time window, it is also obvious that many patients do not receive treatment who would benefit from it and at the same time patients receive treatment subjecting them to risk of intervention such as bleeding without its benefit. This is due to the above-mentioned fact that, in reality, several stroke subpopulations exist. Here, precision medicine accounts for the individual features and will improve outcome by personalizing treatment [25]. Precision medicine utilizes mathematical techniques and available digital data to provide individualized predictions for patients [3]. It relies on the presence of informative data allowing the differentiation of pathology patterns [4]. In stroke, it has been shown that measuring the penumbra through perfusion as a surrogate is one of the most promising approaches [6, 7]. And indeed, perfusion imaging-based selection for treatment beyond the established time windows is an evidenced precision medicine approach in stroke [7]. For the selection process, predictive modelling might be also applicable [5]. A drawback of this approach, however, is the application of perfusion measurement techniques which are potentially harmful through contrast agents, inevitably prolong the imaging time and are problematic to standardize across centers [6, 15]. Similar considerations apply to chronic steno-occlusive disease. These are patients with continuously worsening symptoms of atherosclerosis who have a high likelihood for a future stroke event. In these patients, potentially harmful perfusion imaging techniques

should not be used a priori. Contrast agent-free perfusion imaging methods can be used but are—as mentioned above—hard to standardize. Thus, alternatives are warranted, and mechanistic simulations are promising methods. Here, the relevant (patho)physiological biomarker is not directly measured, but mathematically inferred from conditions recorded through other measurements.

As suggested by our work, in the case of cerebrovascular disease, we can infer information about hemodynamics from the individual vasculature of a given patient. We successfully built a pipeline that can extract the vessel information by segmentation, allows annotation of the vessels and simulates hemodynamic information which we were able to relate to clinical DSC-perfusion imaging through an exploratory visual comparison of DSC perfusion and simulation results. While these results need gold standard validation, they pave the way for further development of techniques that might make the need for perfusion imaging in cerebrovascular disease obsolete for some patients while still providing the necessary information for precision medicine selection of patients for prevention and treatment. Our results are in line with other recent exploratory validation studies [26].

Given that mechanistic simulations work on a priori assumptions about the biological system and perfusion measurements actually record dynamic information, it is unlikely that the information provided by both systems will always be a complete match. This is also evidenced by the low sensitivity and the high specificity. DSC-MRI is sensitive to very small changes in perfusion, whereas a mechanistic simulation is expected to distinguish between relevant categories. Thus, the simulation was not able to pick up on every change noticed by the readers (low sensitivity), but where the simulation found vulnerabilities they were almost always accompanied by corresponding changes in DSC-MRI (high specificity). While the clinical relevance needs to be validated in further studies, our results suggest, that the information might be intersecting enough to allow treatment-relevant predictions on an individualized patient level, and consequently to avoid harmful imaging procedures. Thus, when the question is for example about a general status, i.e., “is there a general vulnerability in the right MCA area for ischemia”, mechanistic simulations might be able to provide this information instead of direct perfusion measurements. Also, since many variants of the CoW exist, the simulation could allow the identification of patients with high-risk for stroke owing to their individual CoW configuration [27].

Another potentially big advantage of mechanistic simulations is the possibility to simulate interventions. In chronic steno-occlusive disease, like carotid stenosis or Moya-Moya disease, potential lumen reopening interventions or EC-IC bypass surgery can be performed. With our solution as presented in this work, it would be feasible to simulate the reopening of a vessel and thus simulate the post-intervention status.

With our framework, it is possible to simulate the response of the vasculature to changes in blood pressure. This can potentially be highly important not only for the determination of areas-at-risk for ischemia, but also to predict the response to interventions and surgery, e.g., blood pressure drops during surgery. This is not possible with perfusion measurements, which can only provide a snap-shot of the status quo. Approximations can be done with acetazolamide challenge measurement [28, 29], but this requires repeated measurements, the application of a drug, and can only be performed

within the physiologically tolerable range. A clear advantage of direct dynamic perfusion measurements, on the other hand, is with high likelihood still the recording of subtle changes and small lesions. Importantly, we thus do not claim that mechanistic modeling might be able to make all perfusion measurements obsolete.

As a limitation of our work, our results are exploratory and hypothesis-generating [30]. While our exploratory validation yielded promising results, further validations are needed. We believe, however, that our results are motivating to boost the translation of the work done in the past on the translational development of mechanistic modeling of hemodynamics into the clinical setting. We implemented a 0-dimensional (D) model of hemodynamics which exploits the similarities of such a network to an electric circuit. Next to these 1D and 3D models exist; for an overview of existing methods see Leguy et al. and Perera et al. [31, 32]. While 1D and 3D models are more suited to model local changes, 0D models are more suited to model the general vasculature, but they can be combined to provide complementary information [33]. Here, our framework builds on existing work, but adds (A) the calculation of the resistance over segments with variable diameters. (B) an easy to code-adjusted implementation of the modified nodal analysis which reduces computational demands and (C) a graphical user interface tailored for inspection and manipulation of the simulation. This facilitates the application of such a framework, which is promising as there is much promise in mechanistic modeling of blood flow and perfusion for clinical applications in cerebrovascular disease. There is, however, a need to personalize these approaches.

Our work has other limitations. First, the simulation values were not compared to a gold standard. However, it is very difficult to derive *individualized* gold standard values for arterial flow. Other noninvasive methods are either non-gold standards themselves, like MRI flow measurement methods, and/or cannot access the complete vasculature, like Doppler-sonography. Intra-operative direct vessel flow measurements recorded during EC–IC bypass surgery might be an option for gold standard measurements—which we will explore in the future—but were not available for the current study. Second, we would like to point out that all relevant pre-processing steps in the pipeline—segmentation, skeletonization, and annotation—were done manually in the pipeline. With the advent of powerful machine learning segmentation methods in recent years, it is very likely that these steps can be automated with sufficient performance. For segmentation, our group has just recently presented deep learning methods to segment the vasculature from structural scans with very high accuracy [25, 34]. The application of deep learning for skeletonization and automated annotation is a current focus of our group. We are thus confident that for future potential applications in the clinical setting simulation results will be obtainable in real-time, at the scanner console, in a few years.

Conclusion

We present the first precision medicine pipeline for cerebrovascular disease that allows annotation of the arterial vasculature derived from structural vessel imaging followed by personalized simulation of brain hemodynamics. This enables further development of precision medicine in stroke using novel biomarkers and might make the application of harmful and complex perfusion methods obsolete for certain use cases.

Table 2 Patient characteristics

Sex	Age (years)	NIHSS	Modified Rankin Scale	Previous cerebrovascular events
Female				
28	Median: 57	0: 47	0: 47	Stroke: 47
Male				
41	Range: 29–82	1–4: 17	1: 10	TIA:12
		5–15: 5	2: 6	
		16–20: 0	3: 3	
		21–42: 0	4: 3	
			5/6: 0	

NIHSS National Institutes of Health Stroke Scale, TIA transient ischemic attack

Methods

Data accessibility

The datasets presented in this article are not readily available because data protection laws prohibit sharing the imaging data used in this study at the current time point. Requests to access the datasets should be directed to ethikkommission@charite.de.

Patients

Sixty-seven patients with steno-occlusive disease from an imaging study of cerebral perfusion in stroke patients (PEGASUS study [8, 9]) were evaluated as examples for the presented framework. Patient characteristics can be found in Table 2. This study was approved by the institutional ethics committee of Charité Universitätsmedizin Berlin and the patients gave written informed consent.

The framework pipeline

The framework pipeline consists of these chronological steps: (a) segmentation, (b) skeletonization, (c) annotation, and (d) simulation. In the following, each of the steps is described in detail.

Segmentation

Structural MRI imaging consisted of time-of-flight (TOF)-MR images. To save time, all images were first pre-segmented using a high-performing published neural net segmentation model [25]. The pre-segmentation model was applied without modifications. All information about the data employed in the training, the architecture, the training regime and the performance can be found in the open-access publication of Livne et al. [25]. The segmentations were then manually corrected to derive ground-truths standards. The segmentations are voxel-based binary representations of the vessel tree for each individual patient. The imaging parameters for TOF-MR-images in the PEGASUS study were: voxel size = (0.5 × 0.5 × 0.7) mm [3]; matrix size: 312 × 384 × 127; TR/TE = 22 ms/3.86 ms; time of acquisition: 3:50 min, flip angle = 18 degrees.

Skeletonization

The segmentations were skeletonized using the DtfSkeletonization module (DtfSkeletonization—MeVisLab documentation) of MeVisLab (website: MeVisLab) [35]. Here, a one voxel skeleton of the vessel midpoints is created with the radius encoded in the voxel value. This skeleton volume is then transferred to the manual annotation module.

Annotation

The annotation module was developed and coded from scratch by the in-house development team.

Within the annotation framework, the transferred skeleton volume is transformed into a Java-based tree structure—the so-called skeleton graph—representing the skeleton as a set of edges containing all necessary geometric information for 3D rendering, and a set of junction vertices. This tree structure is loaded via a RESTful service interface in JSON format into a JavaScript-based web frontend where it is rendered by using the Three.js library as a rotatable and zoomable 3D view.

Within the 3D view, it is possible to select edges and tag them with an item from a list of vessels (artery) descriptors and their anatomic location (visualized by color) interactively. The triple consisting of an edge, a tagged vessel item, and an anatomic location is defined as an annotation. A set of made annotations can be saved in the backend via a RESTful service interface within an appropriate Java presentation. In the final step, based on a skeleton graph and a belonging annotation set, a simulation model serving as the input for our simulation component can be created once the annotation is finished.

Within our framework, it is possible to annotate 3rd order branches (A3, M3, P3). For exemplary patients in our study, annotations were performed until 2nd order branches (A2, M2 and P2), since higher order vessels are unlikely to play a crucial role in stenocclusive disease.

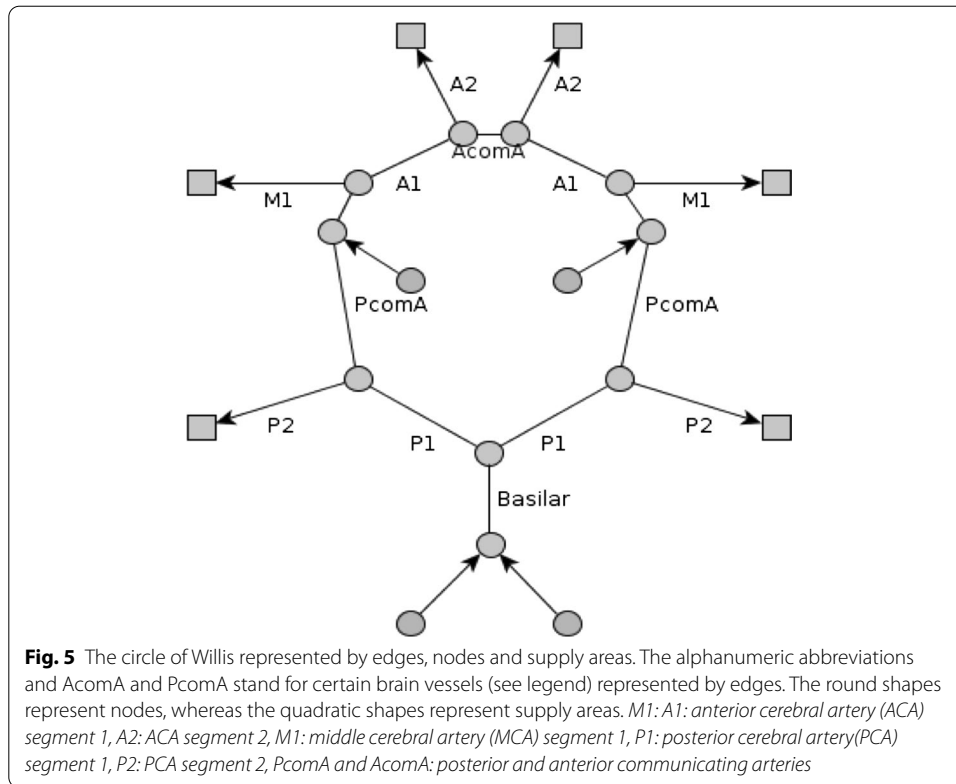
Simulation

The simulation was developed and coded from scratch in-house. Our model describes the cerebral vascular tree by a planar graph, which is given in Fig. 5. The blood flow through the vessel tree is modeled in analogy to electric circuits in a modified nodal analysis.

Edges in the graph represent blood vessels, while nodes represent either supply areas, blood sources or junctions between nodes. Supply areas are marked with square-shaped nodes, initial (round) nodes are blood sources and the rest of the (round) nodes are junctions. The arrows of the edges indicate the flow directions to the supply areas or away from source nodes. The blood flow into the supply areas is provided by the outgoing segments (A2, P2 and M1 or M2) of the circle of Willis.

Modeling the flow and vessel network

The cerebral vessel tree retains an overall Reynolds number allowing to describe the cerebral blood flow in terms of a Newtonian fluid. The arteries are modeled as perfect cylinders and the decrease in blood pressure ΔP along a cerebral artery of length L , with



radius r , blood dynamic viscosity μ for a volumetric flow rate Q is determined according to Hagen–Poiseuille equation:

$$\Delta P = \frac{8 \cdot \mu \cdot L \cdot Q}{\pi r^2}. \tag{1}$$

Hagen–Poiseuille equation is equivalent to Ohm’s law. Therefore, the resistance of an arterial vessel can be defined as:

$$R = \frac{8 \cdot \mu \cdot L}{\pi r^2}. \tag{2}$$

In the presented use case, the vessel diameter is not homogeneous across the whole vessel segment and consequently Eq. 2 is invalid. To account for variable diameters over a segment, the fluid dynamics is applied on infinitely small segments with a constant radius to yield the equation for non-constant radii. The resistance can be therefore derived using the following integral equation:

$$R = \int_L \frac{8\mu}{\pi r(l)^4} dl. \tag{3}$$

In practice, discrete application is used, as a segment is defined by a diameters-vector of length $n-1$ = the number of voxels in the segment. To account for the fact that an antiderivative can only be determined for segments that can be constantly differentiated, we need to approximate the condition of continuity. For this purpose, the

radius $r(l)$ is described by a continuous linear extrapolation function, which connects the radius of a given voxel with the next voxel in a linear fashion to allow the calculation of (3):

$$R_{ext} = \sum_{i=1}^{n-1} \begin{cases} -\frac{8\mu}{3\pi} \left(\frac{r_{i+1}-r_i}{l_{i+1}-l_i}\right)^{-1} (r_{i+1}^{-3} - r_i^{-3}), & r_{i+1} \neq r_i \\ \frac{8\mu}{\pi} (l_{i+1} - l_i) (r_i)^{-4}, & r_{i+1} = r_i \end{cases}, \quad n \in \mathbb{N}, \quad (4)$$

where R_{ext} stands for the extrapolated resistance.

According to the mass flow law, the amount of blood entering a node must equal the amount of blood that leaves a node (see illustration in Fig. 6a). Mathematically this is described according to Kirchhoff's first law as follows:

$$\sum_{i=1}^N Q_{inp,i} = \sum_{j=1}^M Q_{out,j}, \quad (5)$$

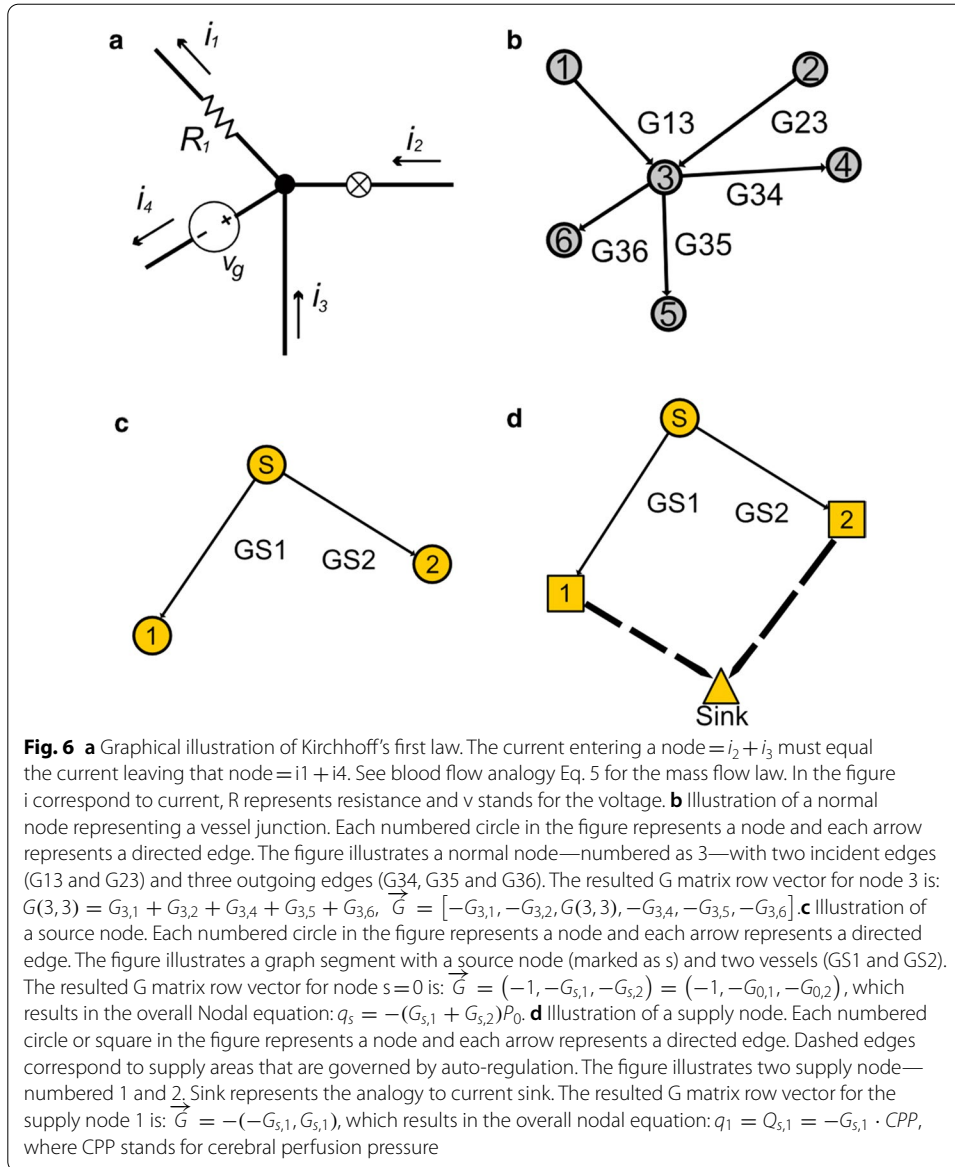
where $Q_{inp,i}$ is the input blood flow from source i to node n , for N input sources, and $Q_{out,j}$ is the output from node n through vessel j , for M outputs.

The mass flow conservation can be then applied using this equation for each node and the resulting set of equations can be solved to yield the pressures. In the presented application, it is applied via a modified nodal analysis (MNA) [21] that incorporates constraints on the system to be driven by a system pressure and ensures constant blood supply to specific regions.

Our network consists of three types of nodes and two types of edges. The primary type of node is one that connects different vessels with each other (i.e., junctions between vessels). The second type of node is a source that provides the system with blood. The third type of node is a supply area, whose resistance depends on the incident pressure and the auto-regulation process described in the next section (see Fig. 5). The types of edges in our network represent normal arterial blood vessels and vessels that connect the supply areas to the sink node. The resistance of the latter is determined via the auto-regulation function given in the next section.

Modeling auto-regulation

The simulated vessel network encompasses the circle of Willis and the larger arterial segments of 1st and 2nd order of the three major brain arteries: anterior-, medial- and posterior cerebral artery (ACA, MCA and PCA). Those are the vasculature segments that are (a) accessible for interventions such as surgery or thrombectomy and (b) their anatomical architecture can be derived from medical imaging. The vascular downstream regions after the above segments including the 3rd order segments represent a network of small arteries, finer small arterioles and the capillary bed that can change their radii in order to decrease or increase the blood supply. This process is called auto-regulation and ensures that the blood supply to the brain remains largely constant within certain limits. Most of the vascular network's resistance originates from these supply areas. Autoregulation was implemented into the simulation framework according to the following equation based on literature values [36]:



$$R_{\text{auto}}(P_{\text{inp}}, P_{\text{out}}) = \begin{cases} \frac{P_{\text{inp}} - P_{\text{out}}}{Q_{\text{supply}}} & \text{if } 150\text{mm Hg} > P_{\text{inp}} > 50\text{mm Hg} \\ R_{\text{min}} = \frac{50\text{mm Hg}}{Q_{\text{supply}}} & \text{if } P_{\text{inp}} < 50\text{mm Hg} \\ R_{\text{max}} = 2.25 * R_{\text{min}} & \text{if } P_{\text{inp}} > 150\text{mm Hg} \end{cases}, \quad (6)$$

where R_{auto} is the autoregulated resistance of the vessel, P_{inp} and P_{out} are the input and output blood pressures of the vessel and Q_{supply} is the blood flow supply to the vessel.

In more detail, the blood flow into the supply areas is provided by the outgoing segments (A2, P2 and M1 or M2) of the circle of Willis, see Fig. 5. The behavior of the supply areas is modeled according to the autoregulation as described in Eq. 6. This means that the peripheral resistance of the supply area adjusts itself such that the blood flow is kept constant for a given pressure gradient ΔP .

Boundary conditions

The model requires a systemic mean arterial pressure (MAP) that drives the flow through the network such that the blood supply to the various supply areas of the brain is constant, while the blood flows and pressures are adjusted accordingly. However, the auto-regulation model is limited to the region between 50 and 150 mmHg. Blood pressure below or above these boundaries indicates a pathological state in which the body cannot maintain the necessary blood pressure to ensure constant blood supply and results in hypoperfusion of the brain tissue. Under these conditions, the simulation indicates that the respective supply areas are not sufficiently supplied with blood. Clinically speaking, these areas are vulnerable to ischemia. In addition to the system pressure a blood flow rate per supply area was provided as P_{in} being the system pressure, while $Q_{s,i}$ for $i = \{1, \dots, N\}$ being the blood supply demand of supply area i as physical quantity.

Algorithmic derivation of the blood supply

The blood supply to the specified brain areas is derived using an adjusted modified nodal analysis (AMNA). The construction of the matrix equations per type of node and edge is detailed in the following. The described system is overdetermined by N equations, where N stands for the number of nodes. The last node is taken as the sink node, with a pressure value of 0. Therefore, the system is described by $N-1$ mass flow equations. These mass flow equations can be written in terms of a matrix:

$$G \cdot \vec{P} = \vec{Q}, \quad (7)$$

where the matrix G represents the network structure and consists of the conductivity (e.g., inverse resistances) of the vessels. \vec{P} is the unknown blood pressure vector and \vec{Q} is the blood flow vector. The AMNA algorithm yields a reduced size of the solution system by removal of known values from the solution vector \vec{Q} . The following section details the determination of the conductivity values and blood flow to construct the matrix equations from a vessel graph as depicted in Fig. 5 and followed by the AMNA implementation.

Junction nodes

The matrix row vector for a junction node is defined as:

$$G(i, j) = \begin{cases} -G_{ij}, & i \neq j \\ \sum_j G_{ij}, & i = j, \end{cases} \quad (8)$$

where the diagonal element is the sum of all G values for all connected edges and the non-diagonal elements are set to be the negative conductivity value of the corresponding edge—or otherwise 0. An exemplary derivation of a junction node vector is described under Fig. 6b. The corresponding element of the \vec{Q} is 0. This represents a flow equation as depicted by Eq. 5.

Source nodes

The matrix row vector for a source node is defined as:

$$G(i,j) = \begin{cases} -G_{i,j}, & i \neq j \\ -1, & i = j \end{cases}. \quad (9)$$

The diagonal element is $G(i,j) = -1$ and all other matrix elements are the negative conductivity values of the (connected) corresponding edges and otherwise 0, similarly to junction nodes. An exemplary derivation of the vector is described in Fig. 6c. The flow into the source node is determined via the supply areas and according to the mass flow can be written as:

$$q_s = -P_0 \cdot \sum G_{s,i}, \quad \text{for Node } i \text{ connected to source } s. \quad (10)$$

For the nodes that are directly connected with the source node, the q vector element is then:

$$q_{\text{connected}} = G_{s,\text{connected}} \cdot P_0. \quad (11)$$

Figure 6c depicts a graph segment with a source node and two connected vessels.

Supply nodes

To model supply areas, the auto-regulation equation is applied (Eq. 6). The matrix row vector for a supply node is defined as:

$$G(i,j) = \begin{cases} -G_{i,j}, & i \neq j \\ \sum_j G_{i,j}, & i = j, j \neq \text{sink} \end{cases}. \quad (12)$$

Similarly to junction nodes, the diagonal element $G(i,j)$ is the conductivity sum of incident edges, however in case the edge is connected to a sink it does not contribute to the sum. An exemplary derivation of the vector is described in Fig. 6d. The q-vector values are:

$$q_i = -Q_{s,i} \text{ for the supply node } i \text{ connected to source } s.$$

$$\text{And } q_{\text{sink}} = -Q_{s,i} \text{ for the sink connected to supply node } i.$$

Figure 6d depicts a case with two supply areas, denoted by the thick dashed lines that are connected to the same sink.

Summary of nodal analysis construction

To summarize, the q-vector has a contribution for a supply area or a source node. If an edge connects a source with a supply area, the corresponding q-vectors will have contributions from both the source and the supply area. Table 3 summarizes the terms for the q-vector elements in dependence of the node type:

$$G_{i,j} = \begin{cases} R_{\text{ext}}^{-1}, & \text{vessel} \\ 0, & \text{supply} \\ R_{\text{auto}}^{-1}, & \text{sink} \end{cases}. \quad (13)$$

Table 3 A: Summary of q vector element derivation in the modified nodal analysis (MNA)

Type of node	Q-vector node	Q-vector incident node
A		
Source	$q_s = -P_0 \cdot \sum G_{s,i}$	$G_{s,connected} P_0$
Supply area	$-Q_{s,js}$	$Q_{s,js}$
Junction	0	0
Element	Value	
B		
$G(i, i)$	$\sum_{i \neq j}^N G_{ij}$	
$G(i, j)$	$-G_{ij}$	

B: Summary of the algorithm for the creation and determination of the G-matrix. The edge conductivity is described by Eq. (13). Here, the edge conductivity between a supply node and the sink is determined by the auto-regulation function

Here vessel refers to a connection of junction node to junction node, supply refers to a connection of junction node to a supply node and sink refers to a connection of a supply node to sink. R_{ext} and R_{auto} are described in formulas 4 and 6, respectively.

Adjusted modified nodal analysis (AMNA)

The AMNA allows to reduce the size of the solution system by removing known values from the solution vector \vec{Q} in consecutive 3 steps. Once the MNA matrix equations are determined, all the values of the system matrix that are known given system pressures are drawn to the right-side solution vector \vec{Q} . This pertains the values associated with the input pressures of the source nodes, i.e., MAP. As a result of this transition the equations associated with the source nodes become zero. In a second step, these redundant rows and columns are then deleted. Finally, the G matrix columns are swapped to yield a diagonal matrix, with the solution vector adjusted accordingly. This process simplifies the solution derivation and therefore accelerates its application.

Sensitivity analysis

In order to predict the effects of changes of a certain variable on the system, in this case the radius of a blood vessel, we performed a sensitivity analysis. Sensitivity analysis quantifies this effect by estimating the partial derivative of a system variable such as the blood pressure in this case, with respect to the radius of a given vessel. This is achieved by application of Newton–Raphson method using Taylor series expansion [37] as follows where—as previously defined—Q is the blood pressure and l is the length of the artery. The formula presented shows the sensitivity analysis—as an example—for the internal carotid artery denoted as int.car.I:

$$\frac{\partial Q_{A1}}{\partial r_{Int.Car.I}} \approx [Q_{A1}(r_{A1}, l_{A1}, \dots, r_{Int.Car.I}, l_{Int.Car.I}, p_{in}) - Q_{A1}(r_{A1}, l_{A1}, \dots, r_{Int.Car.I} + \Delta r_{Int.Car.I}, l_{Int.Car.I}, p_{in})] / \Delta r_{Int.Car.I}. \quad (14)$$

Similar Newton steps can be formulated for all relevant system variables and allow to estimate how the system reacts to changes of certain system variables.

Simulation interface

The simulation interface was implemented as a java application with an integrated graphical user interface (GUI) under the loose-coupling paradigm to ensure that components can be exchanged easily. The key element of the simulation is a 2D projection as a representation of the simulated vessel tree. The interface's main components are the simulation view including areas at risk and pressure view, and the possibility to change blood pressure as a boundary condition.

Perfusion imaging processing

DSC imaging was processed using the pgui software (Version 1.0, Center for functional neuroimaging, Aarhus University). Four arterial input functions were placed contralateral to the stenosis/occlusion in the M2 vessel area and visually assessed for optimal shape [16]. Deconvolution was performed according to the parametric method introduced by Mouridsen et al. [38]. Non-deconvolved time-to-peak, and deconvolved cerebral blood flow (CBF), time-to-maximum (Tmax) and mean-transit-time(MTT) maps were created and assessed in this study.

Comparison of simulation and perfusion imaging results

In 67 patients, we compared simulation and DSC perfusion imaging results. For this purpose, we defined vulnerability for both modalities and determined the detection rate of the simulation to identify vulnerabilities in DSC imaging by sensitivity and specificity.

Vulnerability of brain tissue to ischemia was defined in MR-imaging as a visually rated TTP and/or Tmax increase and/or CBF decrease. For the simulation, vulnerability was defined as a perfusion pressure below 50 mmHg, at normal blood pressure or at a mean-arterial-pressure of 70. We assessed the sensitivity/specificity of the simulation results to detect vulnerability as defined by the visual DSC-analysis. Results were recorded for anterior/middle/posterior cerebral artery regions (left/right).

Acknowledgements

This work was published previously on the MedRxiv pre-print server [39]. We acknowledge Kim Mouridsen and Mikkel Bo Hansen from the Centre for Functionally Integrative Neuroscience (CFIN) from Aarhus University, Denmark, for providing the pgui perfusion software (V1.0) for research purposes.

Authors' contributions

Conceptualization: DF, ML, HL, JS, PV, VIM. Funding acquisition: DF, JS, PV. Project administration: DF, VIM. Data curation: DF, ML, EMA, OUA, JB, VIM. Methodology: DF, ML, HL, JS, PV, VIM. Programming of the frameworks: HL. Writing—original draft: DF, ML, HL, VIM. Writing—review and editing: DF, ML, HL, EMA, OUA, JB, JS, PV, VIM. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work has received funding by the German Federal Ministry of Education and Research through (1) a GO-Bio grant for the research group PREDICTioN2020 (lead: DF), funding from the European Commission via the Horizon 2020 program for PRECISE4Q (No. 777107, lead: DF) and (2) the Centre for Stroke Research Berlin.

Availability of data and materials

Due to privacy laws and regulations the dataset which contains sensitive patient data cannot be made available.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional ethics committee of Charité Universitätsmedizin Berlin and the patients gave written informed consent.

Consent for publication

Not applicable.

Competing interests

The authors report no conflicts of interest.

Author details

¹Charite Lab for Artificial Intelligence in Medicine, Department of Neurosurgery, Charité University Medicine Berlin, Chariteplatz 1, 10115 Berlin, Germany. ²Johanna Etienne Hospital Neuss, Berlin, Germany. ³Centre for Stroke Research Berlin, Charité University Medicine Berlin, Berlin, Germany. ⁴Department of Neurosurgery, Charité University Medicine Berlin, Berlin, Germany. ⁵School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, UK.

Received: 7 January 2021 Accepted: 20 April 2021

Published online: 01 May 2021

References

1. WHO EMRO | Stroke, Cerebrovascular accident | Health topics (2019). <http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html>. Accessed 14 July 2020.
2. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics—2019 Update: a Report From the American Heart Association. *Circulation*. 2019. <https://doi.org/10.1161/CIR.0000000000000659>.
3. Rostanski SK, Marshall RS. Precision medicine for ischemic stroke. *JAMA Neurol*. 2016;73:773–4. <https://doi.org/10.1001/jamaneurol.2016.0087>.
4. Hinman JD, Rost NS, Leung TW, Montaner J, Muir KW, Brown S, et al. Principles of precision medicine in stroke. *J Neurol Neurosurg Psychiatry*. 2017;88:54–61. <https://doi.org/10.1136/jnnp-2016-314587>.
5. Livne M, Boldsen JK, Mikkelsen IK, Fiebach JB, Sobesky J, Mouridsen K. Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke. *Stroke*. 2018;49:912–8. <https://doi.org/10.1161/STROKEAHA.117.019440>.
6. Sobesky J. Refining the mismatch concept in acute stroke: lessons learned from PET and MRI. *J Cereb Blood Flow Metab*. 2012;32:1416–25. <https://doi.org/10.1038/jcbfm.2012.54>.
7. Campbell BCV, Ma H, Ringleb PA, Parsons MW, Churilov L, Bendszus M, et al. Extending thrombolysis to 4.5–9 h and wake-up stroke using perfusion imaging: a systematic review and meta-analysis of individual patient data. *Lancet*. 2019;394:139–47. [https://doi.org/10.1016/S0140-6736\(19\)31053-0](https://doi.org/10.1016/S0140-6736(19)31053-0).
8. Mutke MA, Madai VI, von Samson-Himmelstjerna FC, Zaro Weber O, Revankar GS, Martin SZ, et al. Clinical evaluation of an arterial-spin-labeling product sequence in steno-occlusive disease of the brain. *PLoS ONE*. 2014;9:e87143. <https://doi.org/10.1371/journal.pone.0087143>.
9. Martin SZ, Madai VI, von Samson-Himmelstjerna FC, Mutke MA, Bauer M, Herzog CX, et al. 3D GRASE pulsed arterial spin labeling at multiple inflow times in patients with long arterial transit times: comparison with dynamic susceptibility-weighted contrast-enhanced MRI at 3 Tesla. *J Cereb Blood Flow Metab*. 2015;35:392–401. <https://doi.org/10.1038/jcbfm.2014.200>.
10. Wintermark M, Albers GW, Broderick JP, Demchuk AM, Fiebach JB, Fiehler J, et al. Acute Stroke Imaging Research Roadmap II. *Stroke*. 2013;44:2628–39. <https://doi.org/10.1161/STROKEAHA.113.002015>.
11. Wintermark M, Sesay M, Barbier E, Borbély K, Dillon WP, Eastwood JD, et al. Comparative overview of brain perfusion imaging techniques. *J Neuroradiol J Neuroradiol*. 2005;32:294–314.
12. Khalil AA, Villringer K, Filleböck V, Hu J-Y, Rocco A, Fiebach JB, et al. Non-invasive monitoring of longitudinal changes in cerebral hemodynamics in acute ischemic stroke using BOLD signal delay. *J Cereb Blood Flow Metab*. 2018. <https://doi.org/10.1177/0271678X18803951>.
13. Kamalian S, Lev MH. Stroke imaging. *Radiol Clin*. 2019;57:717–32. <https://doi.org/10.1016/j.rcl.2019.02.001>.
14. Okell TW, Harston GWJ, Chappell MA, Sheerin F, Kennedy J, Jezzard P. Measurement of collateral perfusion in acute stroke: a vessel-encoded arterial spin labeling study. *Sci Rep*. 2019;9:1–10. <https://doi.org/10.1038/s41598-019-44417-7>.
15. Zaharchuk G. Arterial spin label imaging of acute ischemic stroke and transient ischemic attack. *Neuroimaging Clin N Am*. 2011;21:285–301. <https://doi.org/10.1016/j.nic.2011.01.003>.
16. Calamante F. Arterial input function in perfusion MRI: A comprehensive review. *Prog Nucl Magn Reson Spectrosc*. 2013;74:1–32. <https://doi.org/10.1016/j.pnmrs.2013.04.002>.
17. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16:150. <https://doi.org/10.1186/s12916-018-1122-7>.
18. Alastruey J, Parker KH, Peiró J, Byrd SM, Sherwin SJ. Modelling the circle of Willis to assess the effects of anatomical variations and occlusions on cerebral flows. *J Biomech*. 2007;40:1794–805. <https://doi.org/10.1016/j.jbiomech.2006.07.008>.
19. Grinberg L, Anor T, Cheever E, Madsen JR, Karniadakis GE. Simulation of the human intracranial arterial tree. *Philos Trans R Soc Math Phys Eng Sci*. 2009;367:2371–86. <https://doi.org/10.1098/rsta.2008.0307>.
20. Wessels L, Hecht N, Vajkoczy P. Bypass in neurosurgery—indications and techniques. *Neurosurg Rev*. 2019;42:389–93. <https://doi.org/10.1007/s10143-018-0966-9>.
21. Ho C-W, Ruehli A, Brennan P. The modified nodal approach to network analysis. *IEEE Trans Circuits Syst*. 1975;22:504–9. <https://doi.org/10.1109/TCS.1975.1084079>.
22. Frey D, Livne M, Leppin H, Akay EM, Aydin OU, Behland J, et al. Video footage of vessel annotation framework. 2019. Zenodo. <https://doi.org/10.5281/zenodo.3576353>.

23. Marchal G, Beaudouin V, Rioux P, de la Sayette V, Doze FL, Viader F, et al. Prolonged persistence of substantial volumes of potentially viable brain tissue after stroke: a correlative PET-CT Study With Voxel-Based Data Analysis. *Stroke*. 1996;27:599–606. <https://doi.org/10.1161/01.STR.27.4.599>.
24. Powers WJ, Rabinstein AA, Teri A, Adeoye OM, Bambakidis NC, Kyra B, 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:e46–99. <https://doi.org/10.1161/STR.000000000000158>.
25. Livne M, Rieger J, Aydin OU, Taha AA, Akay EM, Kossen T, et al. A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Front Neurosci*. 2019. <https://doi.org/10.3389/fnins.2019.00097>.
26. Helthuis JHG, van Doormaal TPC, Amin-Hanjani S, Du X, Charbel FT, Hillen B, et al. A patient-specific cerebral blood flow model. *J Biomech*. 2020;98:109445. <https://doi.org/10.1016/j.jbiomech.2019.109445>.
27. Pascalau R, Padurean VA, Bartoş D, Bartoş A, Szabo BA. The geometry of the circle of willis anatomical variants as a potential cerebrovascular risk factor. *Turk Neurosurg*; 2018. <https://doi.org/10.5137/1019-5149.JTN.21835-17.3>.
28. Yen Y-F, Field AS, Martin EM, Ari N, Burdette JH, Moody DM, et al. Test-retest reproducibility of quantitative CBF measurements using fAIR perfusion MRI and acetazolamide challenge. *Magn Reson Med*. 2002;47:921–8. <https://doi.org/10.1002/mrm.10140>.
29. Ma J, Mehrkens JH, Holtmannspoetter M, Linke R, Schmid-Elsaesser R, Steiger H-J, et al. Perfusion MRI before and after acetazolamide administration for assessment of cerebrovascular reserve capacity in patients with symptomatic internal carotid artery (ICA) occlusion: comparison with 99mTc-ECD SPECT. *Neuroradiology*. 2007;49:317–26. <https://doi.org/10.1007/s00234-006-0193-x>.
30. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation. *PLoS Biol*. 2014;12:e1001863. <https://doi.org/10.1371/journal.pbio.1001863>.
31. Leguy C. "Mathematical and Computational Modelling of Blood Pressure and Flow," in *Cardiovascular Computing—Methodologies and Clinical Applications Series in BioEngineering*, eds. S. Golemati and K. S. Nikita (Singapore: Springer Singapore), 2019; 231–246. doi: https://doi.org/10.1007/978-981-10-5092-3_11.
32. Perera K. Literature Review on Methods of Modeling the Cerebral Network and the Circle of Willis. <http://mars.gmu.edu/handle/1920/11429>. Accessed 3 July 2019.
33. Chau N, Ho H. A Hybrid 0D–1D Model for Cerebral Circulation and Cerebral Arteries. in *Computational Biomechanics for Medicine*. In: Nash MP, Nielsen PMF, Wittek A, Miller K, Joldes GR, eds. (Springer International Publishing), 2020; 99–110.
34. Hilbert A, Madai VI, Akay EM, Aydin OU, Behland J, Sobesky J, et al. BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. *Front Artif Intell*. 2020. <https://doi.org/10.3389/frai.2020.552258>.
35. DtfSkeletonization — MeVisLab documentation Available at: <https://mevislabdownloads.mevis.de/docs/2.5/MeVisLab/Standard/Documentation/Publish/ModuleReference/DtfSkeletonization.html>. Accessed 5 Sept 2020.
36. Lang EW, Mudaliar Y, Lagopoulos J, Dorsch N, Yam A, Griffith J, et al. A review of cerebral autoregulation: Assessment and measurements. *Anaesth.: Australas*; 2005. p. 161.
37. Gopalakrishna HS, Greimann LF. Newton-raphson procedure for the sensitivity analysis of nonlinear structural behavior. *Comput Struct*. 1988;30:1263–73. [https://doi.org/10.1016/0045-7949\(88\)90191-5](https://doi.org/10.1016/0045-7949(88)90191-5).
38. Mouridsen K, Hansen MB, Østergaard L, Jespersen SN. Reliable estimation of capillary transit time distributions using DSC-MRI. *J Cereb Blood Flow Metab*. 2014;34:1511–21. <https://doi.org/10.1038/jcbfm.2014.111>.
39. Frey, D., Livne, M., Leppin, H., Akay, E. M., Aydin, O. U., Behland, J., et al. (2020). A Precision Medicine Framework for Personalized Simulation of Hemodynamics in Cerebrovascular Disease. *medRxiv*, 2020.01.28.20019190. doi:<https://doi.org/10.1101/2020.01.28.20019190>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



2.3 Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks³⁴

Für die Realisierung datengetriebener Entscheidungsunterstützungssysteme sind für das Training und Testing der zugrundeliegenden Modelle große Datenmengen erforderlich. In der medizinischen Grundlagenforschung besteht – vor allem im Hinblick auf Bildungsdaten – häufig die Herausforderung, dass suffiziente Datenmengen aufgrund von Vorgaben zur Datensicherheit, fehlendem Einverständnis und protektiven Gesetzen, Richtlinien und Regelungen nicht zugänglich sind. Die in dieser Arbeit vorgestellte Datensynthese von MRT-Bilddaten mittels Generative Adversarial Networks (GANs) kann eine Lösung für dieses Problem der Datenknappheit darstellen.

Die Anonymisierung von persönlichen Patientendaten ist entscheidend für den Schutz der Privatsphäre in der medizinischen Forschung. Insbesondere für die Analyse und Erfassung großer Datensätze in der medizinischen Bildanalyse stellt dies eine große Herausforderung dar. In der MRT-Diagnostik des Gehirns ermöglicht die patientenindividuelle einzigartige Struktur des Gehirns prinzipiell eine Re-Identifizierung des individuellen Patienten. Aus diesem Grund sind besondere sogenannte nicht-konventionelle Anonymisierungsmethoden zu entwickeln. Insbesondere Generative Adversarial Networks (GANs) haben das Potenzial, anonyme Bilder zu generieren und gleichzeitig das Ursprungsbild nicht so weit zu verfälschen, dass die hierin enthaltenen prädiktiven Eigenschaften verloren gehen.

Methodisch haben wir für die Generierung von Hirngefäßen die folgenden GANs auf Time-of-Flight (TOF)-Magnetresonanzangiographie (MRA)-Patches trainiert: 1) Deep Convolutional GAN, 2) Wasserstein-GAN mit Gradient Penalty (WGAN-GP) und 3) WGAN-GP mit spektraler Normalisierung (WGAN-GP-SN). Die von jedem GAN generierten Bild-Labels wurden zum Training eines U-Nets für die Segmentierung verwendet und an realen Daten getestet. Darüber hinaus haben wir die generierten synthetischen Patches mittels Transfer Learning auf einem zweiten Datensatz getestet. Für 15 Patienten haben wir die Modellleistung auf realen Daten mit und ohne Vortraining bewertet. Die Leistung aller Modelle wurde durch den Dice Similarity Coefficient (DSC) und das 95. Perzentil der Hausdorff-Distanz (95HD) bewertet. Beim Vergleich der drei GANs zeigte das U-Net, das auf synthetischen, vom WGAN-GP-SN generierten Daten trainiert wurde, die höchste Performanz (DSC/95HD 0,85/30,00), gefolgt von dem auf realen Daten trainierten U-Net (0,89/26,57).

Damit konnte mit dieser Arbeit gezeigt werden, dass synthetische Bild-Label-Paare eine gute Leistung bei der Gefäßsegmentierung aufweisen und synthetische Patches in einem Transfer-Learning-Ansatz mit unabhängigen Daten verwendet werden können. Insgesamt kann der entwickelte GAN-Ansatz demzufolge einen Weg darstellen, das Problem der Datenknappheit in der Forschung mit medizinischen Bilddaten zu lösen.

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>

2.4 Toward Sharing Brain Images: Differentially Private TOF-MRA Images With Segmentation Labels Using Generative Adversarial Networks³⁵

Für die Entwicklung von Deep-Learning basierten Ansätzen ist die Nutzung gelabelter Daten erforderlich. In der medizinischen Bildgebung ist das Teilen und die Weitergabe von großen Datensätzen aufgrund von Datenschutzbestimmungen oft nicht möglich. Während eine komplette Anonymisierung der Daten eine Lösung wäre, hat sich das Problem gestellt, dass viele der Standardtechniken der Anonymisierung reversibel sind, damit Patienteninformation wieder zugänglich gemacht werden können und dementsprechend keinen absoluten Schutz bieten. Die künstliche Synthese von Daten unter Verwendung eines Generative Adversarial Network (GAN) mit differenziellen Datenschutzgarantien könnten eine Lösung sein, um die Privatsphäre des Patienten zu gewährleisten und gleichzeitig die prädiktiven Eigenschaften der Daten zu erhalten.

In der hier vorgestellten Arbeit haben wir ein Wasserstein-GAN (WGAN) mit und ohne differenzielle Datenschutzgarantien implementiert, um Time-of-Flight Magnetresonanzangiographie (TOF-MRA) Bildelemente für die Segmentierung von Hirngefäßen zu erzeugen. Die synthetisierten Bild-Label-Paare wurden verwendet, um ein U-Net zu trainieren, das in Bezug auf die Segmentierungsleistung auf realen Patientenbildern aus zwei verschiedenen Datensätzen evaluiert wurde (downstream task). Zusätzlich wurde die Fréchet Inception Distance (FID) zwischen den generierten Bildern und den realen Bildern berechnet, um deren Ähnlichkeit zu bewerten.

Während der Bewertung mit dem U-Net und der FID untersuchten wir die Auswirkungen verschiedener Datenschutzstufen, die durch den Parameter ϵ (Epsilon) dargestellt wurden. Hierbei hat sich gezeigt, dass je stärker die Anonymisierung - und damit der Datenschutz - erhöht wurde die Segmentierungsleistung und die Ähnlichkeit mit den echten Patientenbildern in Bezug auf den FID sanken. Als Ergebnis erreichte das beste Segmentierungsmodell, das auf synthetischen und privaten Daten trainiert wurde, einen Dice Similarity Coefficient (DSC) von 0,75 für $\epsilon = 7,4$ im Vergleich zu 0,84 für $\epsilon = \infty$ in einem Paradigma zur Segmentierung von Hirngefäßen (DSC von 0,69 bzw. 0,88 auf dem zweiten Testsatz).

Wir ermittelten einen Schwellenwert von $\epsilon < 5$, bei dem die Leistung (DSC $< 0,61$) instabil und für die Nutzung unzureichend niedrig wurde. Die unter diesen strengen Datenschutzgarantien synthetisierten annotierten TOF-MRA-Bilder behielten die für die Segmentierung der Hirngefäße erforderlichen prädiktiven Eigenschaften bei.

Die hier präsentierten Ergebnisse stellen einen entscheidenden Schritt für das Teilen und die Nutzung von Daten unter Wahrung der Privatsphäre in der medizinischen Bildgebung dar. Für ein Anwendung auf andere Bildgebungsmodalitäten und für die Validierung der Ergebnisse sind weitere Studien erforderlich.



Toward Sharing Brain Images: Differentially Private TOF-MRA Images With Segmentation Labels Using Generative Adversarial Networks

Tabea Kossen^{1,2*}, Manuel A. Hirzel¹, Vince I. Madai^{1,3,4}, Franziska Boenisch⁵, Anja Hennemuth^{2,6,7}, Kristian Hildebrand⁸, Sebastian Pokutta^{9,10}, Kartikey Sharma⁹, Adam Hilbert¹, Jan Sobesky^{11,12}, Ivana Galinovic¹², Ahmed A. Khalil^{12,13,14}, Jochen B. Fiebach¹² and Dietmar Frey¹

¹ CLAIM-Charité Lab for AI in Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany, ² Department of Computer Engineering and Microelectronics, Computer Vision & Remote Sensing, Technical University Berlin, Berlin, Germany, ³ QUEST Center for Responsible Research, Berlin Institute of Health (BIH), Charité-Universitätsmedizin Berlin, Berlin, Germany, ⁴ Faculty of Computing, Engineering and the Built Environment, School of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom, ⁵ Fraunhofer AISEC, Berlin, Germany, ⁶ Institute for Imaging Science and Computational Modelling in Cardiovascular Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany, ⁷ Fraunhofer MEVIS, Bremen, Germany, ⁸ Department VI Computer Science and Media, Berlin University of Applied Sciences and Technology, Berlin, Germany, ⁹ Department for AI in Society, Science, and Technology, Zuse Institute Berlin, Berlin, Germany, ¹⁰ Institute of Mathematics, Technical University Berlin, Berlin, Germany, ¹¹ Johanna-Etienne-Hospital, Neuss, Germany, ¹² Centre for Stroke Research Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany, ¹³ Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, ¹⁴ Mind, Brain, Body Institute, Berlin School of Mind and Brain, Humboldt-Universität Berlin, Berlin, Germany

OPEN ACCESS

Edited by:

Naimul Khan,
Ryerson University, Canada

Reviewed by:

Alessandro Bria,
University of Cassino, Italy
Zeeshan Ahmad,
Ryerson University, Canada

*Correspondence:

Tabea Kossen
tabea.kossen@charite.de

Specialty section:

This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Artificial Intelligence

Received: 12 November 2021

Accepted: 31 March 2022

Published: 02 May 2022

Citation:

Kossen T, Hirzel MA, Madai VI, Boenisch F, Hennemuth A, Hildebrand K, Pokutta S, Sharma K, Hilbert A, Sobesky J, Galinovic I, Khalil AA, Fiebach JB and Frey D (2022) Toward Sharing Brain Images: Differentially Private TOF-MRA Images With Segmentation Labels Using Generative Adversarial Networks. *Front. Artif. Intell.* 5:813842. doi: 10.3389/frai.2022.813842

Sharing labeled data is crucial to acquire large datasets for various Deep Learning applications. In medical imaging, this is often not feasible due to privacy regulations. Whereas anonymization would be a solution, standard techniques have been shown to be partially reversible. Here, synthetic data using a Generative Adversarial Network (GAN) with differential privacy guarantees could be a solution to ensure the patient's privacy while maintaining the predictive properties of the data. In this study, we implemented a Wasserstein GAN (WGAN) with and without differential privacy guarantees to generate privacy-preserving labeled Time-of-Flight Magnetic Resonance Angiography (TOF-MRA) image patches for brain vessel segmentation. The synthesized image-label pairs were used to train a U-net which was evaluated in terms of the segmentation performance on real patient images from two different datasets. Additionally, the Fréchet Inception Distance (FID) was calculated between the generated images and the real images to assess their similarity. During the evaluation using the U-Net and the FID, we explored the effect of different levels of privacy which was represented by the parameter ϵ . With stricter privacy guarantees, the segmentation performance and the similarity to the real patient images in terms of FID decreased. Our best segmentation model, trained on synthetic and private data, achieved a Dice Similarity Coefficient (DSC) of 0.75 for $\epsilon = 7.4$ compared to 0.84 for $\epsilon = \infty$ in a brain vessel segmentation paradigm (DSC of 0.69 and 0.88 on the second test set, respectively). We identified a threshold of $\epsilon < 5$ for which the

performance (DSC < 0.61) became unstable and not usable. Our synthesized labeled TOF-MRA images with strict privacy guarantees retained predictive properties necessary for segmenting the brain vessels. Although further research is warranted regarding generalizability to other imaging modalities and performance improvement, our results mark an encouraging first step for privacy-preserving data sharing in medical imaging.

Keywords: brain vessel segmentation, differential privacy, Generative Adversarial Networks, neuroimaging, privacy preservation

1. INTRODUCTION

Deep Learning techniques are on the rise in many neuroimaging applications (Lundervold and Lundervold, 2019; Zhu et al., 2019; Hilbert et al., 2020). While showing great potential, they also demand large amounts of data. In medical imaging, data is often limited and medical experts are often needed to manually label the images (Willemink et al., 2020). Thus, large datasets are difficult to acquire. One potential solution would be data sharing. For this, true anonymization, i.e. verifying that no identifying information is leaked, is essential to sustain the patient's privacy which poses a big challenge, especially for neuroimaging (Bannier et al., 2021). For example, face-recognition software has recently identified individuals on medical images (Schwarz et al., 2019) and even face removal techniques can be partially reversed (Abramian and Eklund, 2019). Besides that, the brain itself has a unique structure and cortical foldings can be utilized to identify individuals even in the developing stage (Duan et al., 2020). Consequently, it is highly challenging to truly anonymize brain scans without risking re-identification. A promising remedy is the generation of synthetic data.

For this purpose, Generative Adversarial Networks (GANs) have gained a lot of attention in the past years (Yi et al., 2019). This also holds true for the neuroimaging domain. Here, GANs have shown promising results for synthesized images for different types of imaging (Bowles et al., 2018; Foroozandeh and Eklund, 2020; Kossen et al., 2021) as well as for other medical problems such as segmentation (Cirillo et al., 2020). To ensure the privacy of the training data, GANs can be combined with differential privacy (Xie et al., 2018). Differential privacy is a mathematical framework that provides an upper bound on individual privacy leakage (Dwork, 2008). This way the maximum privacy leakage for every individual in the training data can be quantified. There are extensive studies about GANs with differential privacy for synthesizing natural images and tabular medical data (Xie et al., 2018; Torkzadehmahani et al., 2019; Xu et al., 2019; Yoon et al., 2019, 2020). Recently, Cheng et al. (2021) did a comprehensive study about synthetic images and classification fairness with a varying amount of privacy on various types of imaging data. Among them were also 2D medical datasets such as chest x-rays and melanoma images. Few other studies generated chest x-rays with privacy guarantees as well (Nguyen et al., 2021; Zhang et al., 2021). However, to date, no study has investigated whether 2D synthesized data using a GAN with differential privacy can be utilized for a 3D medical application. Additionally, to the best of our knowledge, GANs with differential privacy have

neither been used to synthesize labels for medical images nor the neuroimaging domain yet.

In this study, we utilized a Wasserstein GAN (WGAN) with and without differential privacy guarantees to synthesize anonymously and labeled 2D Time-of-Flight Magnetic Resonance Angiography (TOF-MRA) image patches for brain vessel segmentation. The generated labeled image patches were evaluated in terms of the segmentation performance by training a U-Net and in terms of image quality using the Fréchet Inception Distance (FID). The trained U-Net was further tested on a second dataset. Overall, we investigated the effect of different levels of privacy. Additionally, we visualized generated images with and without privacy together with the real patient images using t-distributed stochastic neighbor embedding (t-SNE).

In summary, our contributions are:

1. To the best of our knowledge, we are the first to synthesize images with differential privacy guarantees in the neuroimaging domain.
2. We also generate the corresponding segmentation labels to evaluate the image-label pairs in an end-to-end brain vessel segmentation paradigm on 3D medical data for different levels of privacy.
3. For evaluation, we compare the distances between the generated data and both the training and test data to investigate the similarity of the synthesized to the original data.
4. We visualize our generated images with and without differential privacy and the original data using t-SNE.

2. RELATED STUDY

For the synthesis of medical images, deep generative models have demonstrated promising results. Among them, especially GANs and variational autoencoders (VAE) have shown good performance in tasks such as data augmentation (Bowles et al., 2018), image-to-image translations (Isola et al., 2018), or reconstruction (Tudosiu et al., 2020). For the purpose of synthesizing privacy-preserving images, VAE has two disadvantages compared to GANs: First, they produce blurrier images (Wang et al., 2020), and second, the training images are directly fed into the network which makes them more vulnerable to membership inference attacks (Chen et al., 2020).

Hence, in this context, GAN architectures with differential privacy have been used in many previous studies to synthesize non-medical images (Xie et al., 2018; Torkzadehmahani et al.,

2019; Xu et al., 2019) and medical tabular data (Yoon et al., 2019, 2020). However, only few studies have applied GANs with differential privacy to medical images. Additionally, these were restricted to chest x-rays (Cheng et al., 2021; Nguyen et al., 2021; Zhang et al., 2021). So far in the neuroimaging domain, the application of GANs remained without differential privacy (Bowles et al., 2018; Foroozandeh and Eklund, 2020; Kossen et al., 2021).

In the present study, we propose a GAN architecture with differential privacy in the neuroimaging domain. Along with our synthesized images, we generate the segmentation labels for testing our differentially private patches in an end-to-end brain vessel segmentation paradigm.

3. MATERIALS AND METHODS

3.1. Data

In total, 131 patients with cerebrovascular disease from the PEGASUS study (N = 66) and the 1000Plus study (N = 65) were utilized in this study. All patients gave their written informed consent and the studies have been authorized by the ethical review committee of Charité–Universitätsmedizin Berlin. More details on both datasets can be found in Mutke et al. (2014) for the PEGASUS study and Hotter et al. (2009) for the 1000Plus study.

The brain scans were conducted on a clinical 3T whole-body system (Magnetom Trio, Siemens Healthcare, Erlangen, Germany) utilizing a 12-channel receive radiofrequency coil (Siemens Healthcare) for head imaging. For both studies the parameters were: voxel size = (0.5 x 0.5 x 0.7) mm³; matrix size: 312 x 384 x 127; TR/TE = 22 ms/3.86 ms; acquisition time: 3:50 min, flip angle = 18°.

The PEGASUS dataset was split into a training (41 patients), validation (11 patients), and test (14 patients) set. The training set was utilized for training the GANs (refer to **Figure 1**), whereas the validation and test set were utilized for the parameter selection of the U-Net and assessing the generalizable performance of the U-Net, respectively. Additionally, the 65 patients from the 1000Plus dataset were used as a second test set.

For each patient of the training set 1,000 2D image patches and corresponding segmentation masks of size 96x96 were extracted. This patch size has been shown to be the most suitable patch size for Wasserstein based GAN architectures for this use case (Kossen et al., 2021). Due to the overemphasis of background compared to brain vessels, 500 patches showing a vessel in the center were extracted. The remaining 500 patches were extracted randomly. It was verified that all patches were only selected at most once.

3.2. Differential Privacy

To account for the level of privacy of the generated data and provide theoretical privacy guarantees, differential privacy was implemented (Dwork, 2008). A randomized algorithm $f : d \rightarrow R$ satisfies (ϵ, δ) -differential privacy if for any two databases $d_1, d_2 \in d$ that differ from each other by a single sample, the following holds:

$$\Pr[f(d_1) \in S] \leq \exp(\epsilon) * \Pr[f(d_2) \in S] + \delta \tag{1}$$

where $f(d_1)$ and $f(d_2)$ denote the output of f and \Pr the probabilities and with $S \subset R$. δ is the probability that the value of ϵ holds true. With a probability of $1 - \delta$ this equation is equivalent to:

$$\log \left(\frac{\Pr[f(d_1) \in S]}{\Pr[f(d_2) \in S]} \right) \leq \epsilon. \tag{2}$$

Thus, differential privacy holds true if the algorithm’s output for d_1 and d_2 is very similar to each other. In other words, one sample should not have a big impact on the algorithm’s output. This way the privacy of each possible datapoint is preserved. The maximal deviation between the outputs is given by $\exp(\epsilon)$. In this way, ϵ can quantify the level of privacy with small values of ϵ indicating stricter privacy guarantees.

Mironov (2017) proposed Rényi differential privacy, a natural relaxation of differential privacy built upon Rényi divergence. Rényi divergence of order $\alpha > 1$ of two probability distributions P and Q is defined as:

$$D_\alpha(P||Q) := \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha, \tag{3}$$

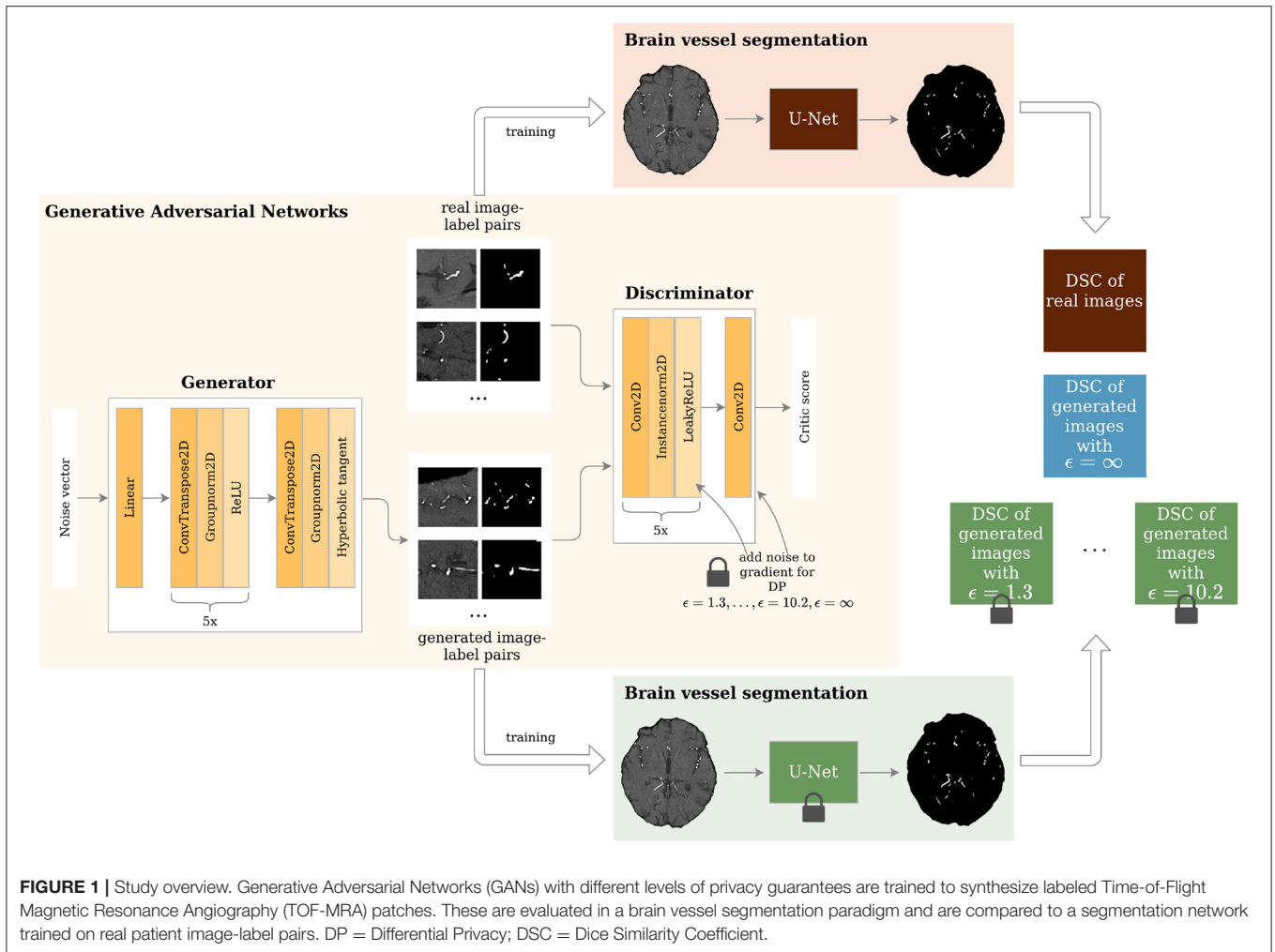
where $P(x)$ is the probability density of P at point x . A randomized algorithm $f : d \rightarrow S$ is (α, ϵ) -Rényi differentially private for any adjacent $d_1, d_2 \in d$ if the Rényi divergence D_α is not larger than ϵ :

$$D_\alpha(f(d_1)||f(d_2)) \leq \epsilon. \tag{4}$$

The advantage of Rényi differential privacy is that it provides a tight composition for Gaussian mechanisms while preserving essential properties of differential privacy. This means that (α, ϵ) -Rényi differential privacy for composed mechanisms add up: the composition of $f(d_1)$ satisfying (α, ϵ_1) -Rényi differential privacy and $f(d_2)$ satisfying (α, ϵ_2) -Rényi differential privacy satisfies $(\alpha, \epsilon_1 + \epsilon_2)$ -Rényi differential privacy. Moreover, (α, ϵ) -Rényi differential privacy has been shown to provide a tighter bound on the privacy budget of compositions compared to (ϵ, δ) -differential privacy (Mironov, 2017). (α, ϵ) -Rényi differential privacy can also be translated back into (ϵ, δ) -differential privacy. Balle et al. (2019) has proven that (α, ϵ) -Rényi differential privacy also satisfies (ϵ', δ) -differential privacy for any $0 < \delta < 1$. According to Balle et al. (2019) ϵ' is then defined as:

$$\epsilon' = \epsilon + \log \frac{\alpha - 1}{\alpha} - \frac{\log \delta + \log \alpha}{\alpha - 1}. \tag{5}$$

The most data sensitive part when training the proposed GAN architecture is the gradient update of the discriminator after training samples are presented. For that, the differentially private stochastic gradient descent algorithm proposed by Abadi et al. (2016) can be utilized. Here, differential privacy was implemented by clipping these gradients and adding Gaussian noise to avoid the memorization of single samples. Additionally, Rényi differential privacy was then used to analyze the privacy guarantees. In the last step, (α, ϵ) -Rényi differential privacy is translated back to (ϵ, δ) -differential privacy. The parameter



δ is typically chosen to be the inverse of the dataset size (Torkzadehmahani et al., 2019). Thus, throughout this study, it was set to $1/41,000 = 2.44e - 5$.

3.3. Network Architecture

The GAN architecture was based on the WGAN by Arjovsky et al. (2017) and extended by inserting different amounts of noise into the gradients of the discriminator in the training process for differential privacy. Two neural networks were trained: the generator G and the discriminator D . The generator synthesized data samples that were then assessed with respect to their realness by the critic or discriminator. The discriminator was fed both real and synthesized data and assigned a critic score for each sample. The score of the synthetic data x_{gen} was used to train the generator. For the generator the overall training loss was:

$$\text{loss}_G = -D(x_{gen}). \tag{6}$$

This way the generator aimed to maximize the realness of the generated samples. In contrast to that, the discriminator intended to minimize the scores for generated samples x_{gen} and maximize

them for patient samples x_{real} :

$$\text{loss}_D = D(x_{gen}) - D(x_{real}) \tag{7}$$

To enforce a Lipschitz constraint and, thus, put a bound on the gradients, the discriminator's weights were clipped after each backpropagation step. This is a simple way to stabilize the training (Arjovsky et al., 2017).

The architecture of the generator and discriminator is shown in **Figure 1**. The generator took a noise vector sampled from a Gaussian distribution of size 128 as input. This was then fed through 1 linear layer and 6 upsampling convolutional layers as shown in **Figure 1**. The generator outputs 2 96 x 96 images - 1 channel for the image and 1 for the segmentation label. The discriminator's input was 2 images: either the real patient image-label pair or the generated one. These were then fed through 6 layers of downsampling convolutional layers as depicted in **Figure 1**. The slope of the LeakyReLU activation was 0.2.

The GANs were implemented in PyTorch 1.8.1 using the library opacus 0.14.0 for the differential privacy guarantees. Our

code was built upon the official GAN example by opacus¹ and is publicly available². The learning rate for both discriminator and generator was 0.00005 using the RMSprop optimizer. The kernel size was 4 with strides of 2. In each epoch, the discriminator was updated 5 times. The network was trained for 50 epochs. To randomly sample the training images, the UniformWithReplacementSampler from the opacus package was used. The sampling rate was the batch size of 32 divided by the number of samples (41,000). The clipping parameter for the WGAN was set to 0.01 and the clipping parameter for the differential privacy was 1. In total, 8 different GANs were trained with varying values of ϵ (noise multiplier was set to $\{\infty, 2, 1.5, 1.2, 1, 0.8, 0.725, 0.65\}$). Each GAN trained with additional noise was trained 5 times for robust results.

All hyperparameters mentioned in the last paragraph were the result of a tuning process and all models were trained on a Tesla V100. The training time of one GAN including evaluation took ~ 1.4 days.

3.4. Performance Evaluation

Among the many metrics to evaluate synthetic data (Yi et al., 2019), we selected three to estimate the quality of our synthesized images. First, we evaluated our synthesized image-label pairs by visual inspection, and second, using the downstream task of segmentation as suggested by Yi et al. (2019). Additionally, we compared the images using the FID as proposed in previous studies (Haarburger et al., 2019; Coyner et al., 2022).

The generated image-label pairs were evaluated by a U-Net for brain vessel segmentation adapted from Livne et al. (2019). After training the GANs, 41,000 image-label pairs were generated. These were used to train 8 U-Net with different hyperparameter settings varying in learning rates, dropout, and classical data augmentation. The best U-Net was then selected based on the best Dice Similarity Coefficient (DSC) on the validation set that included real patient images. The final performance was then evaluated in terms of DSC and balanced average Hausdorff distance (bAHD) on the test set. The DSC that evaluated the segmented voxels is defined as:

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (8)$$

where TP are the true positives, FP are the false positives, and FN are the false negatives. As the DSC quantifies the overlap of the ground truth and prediction scaled by the total number of voxels in ground truth and prediction, it is a robust performance measure for imbalanced segmentations, i.e., images contain more background than segmented area. The bAHD is a newly proposed metric for evaluating segmentations (Aydin et al., 2021):

$$bAHD = \left(\frac{1}{N_G} \sum_{g \in G} \min_{s \in S} d(g, s) + \frac{1}{N_G} \sum_{s \in S} \min_{g \in G} d(s, g) \right) / 2 \quad (9)$$

where N_G is the number of ground truth voxels, G is the set of voxels belonging to the ground truth, and S is the set of voxels

of the predicted segmentation. In other words, the bAHD is the average of the directed Hausdorff distance from the ground truth to the segmentation and the directed Hausdorff distance from the segmentation to the ground truth both scaled by the number of ground truth voxels.

Additionally, the DSC and bAHD of the U-Net models were assessed on the 1000Plus dataset. The GAN and U-Nets were implemented in an end-to-end pipeline. To calculate both DSC and bAHD, we used the EvaluateSegmentation tool by Taha and Hanbury (2015).

As an additional metric, the image quality was measured by the FID (Heusel et al., 2018). The FID is a distance that measures the similarity between images by comparing the activations of a pre-trained Inception-v3 network. Here, the difference between the activations in the pool3 layer of the generated images in contrast to the real images is measured.

$$FID = \|\mu_{real} - \mu_{gen}\|^2 + \text{Tr} \left(\sigma_{real} + \sigma_{gen} - 2(\sigma_{real}\sigma_{gen})^{1/2} \right) \quad (10)$$

with $\mathcal{N}(\mu_{real}, \sigma_{real})$ and $\mathcal{N}(\mu_{gen}, \sigma_{gen})$ as the distributions of the features of the pool3 layer of real and synthesized data, respectively.

To explore to which degree the generated images reproduced the training set, the FID between the synthetic data and both the training and test data was calculated and compared for different levels of privacy.

Finally, we measured the similarity between the images synthesized by the GANs to check whether a model suffered from mode collapse. For each model, we generated 1,000 images and calculated the Structural Similarity Index Measure (SSIM) between them and averaged the values. We repeated this analysis for all 5 runs for each ϵ value, for the model with $\epsilon = \infty$ and the real images. The SSIM between two images x and y is defined as a product of luminance, contrast, and structure according to Wang et al. (2004):

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (11)$$

where μ_x is the average of x , σ_x is the variance, and σ_{xy} is the covariance of x and y . $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ are for stabilization with L being the dynamic range of the pixel values and $k_1 \ll 1$ and $k_2 \ll 1$ small constants.

3.5. Visualization Using t-SNE

Finally, the generated images with and without differential privacy and the real patient images were visualized using a t-SNE (Maaten and Hinton, 2008). t-SNE is an approach to reducing dimensionality while preserving the structure of the high dimensional data points. First, all data points are embedded into a SNE which computed the pairwise similarities utilizing conditional probabilities. For points x_i and x_j the conditional probability $p_{j|i}$ of x_i choosing x_j as its neighbor is defined as

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \quad (12)$$

¹<https://github.com/pytorch/opacus/blob/master/examples/dcgan.py>

²<https://github.com/prediction2020/Labeled-TOF-MRA-with-DP>

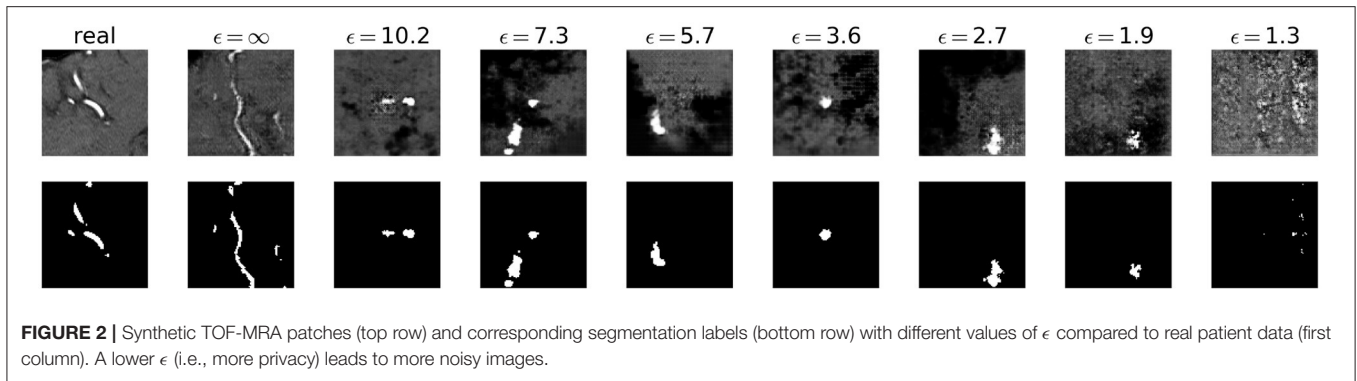


FIGURE 2 | Synthetic TOF-MRA patches (top row) and corresponding segmentation labels (bottom row) with different values of ϵ compared to real patient data (first column). A lower ϵ (i.e., more privacy) leads to more noisy images.

and the symmetrized similarity as:

$$p_{ij} = \frac{p_{ji} + p_{ij}}{2N} \quad (13)$$

with N being the dimensionality of the data. Then the algorithm aims to learn a lower dimensional representation of the similarities. In order to get distinct clusters and avoid overcrowding, a Student's t distribution that reflects the similarities p_{ji} is used (Maaten and Hinton, 2008):

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq m} (1 + \|y_k - y_m\|^2)^{-1}} \quad (14)$$

Starting from random initialization, the locations of the points in the lower dimensional space y_i are shifted so that a cost function was minimized using a gradient descent method. Instead of the Kullback-Leibler divergence, we here chose the Wasserstein metric due to its success in GAN applications (Arjovsky et al., 2017).

T-distributed stochastic neighbor embedding was implemented using the sklearn package (Pedregosa et al., 2011). The perplexity parameter reflecting the density of the data distribution was chosen to be 30 which is in the suggested range by Maaten and Hinton (2008). The images of the best performing GAN with and without differential privacy, as well as the real images were projected, onto 2 dimensions for visualization purposes.

4. RESULTS

Visually, the synthetic image-label pairs appeared noisier with decreasing ϵ , i.e., with stricter privacy guarantees (Figure 2). Differentially private images with $\epsilon = 1.3$ show almost only noise. The visual results corresponded to the segmentation performance when training a U-Net on the generated image-label pairs with different values of ϵ (Figure 3). In Figure 3A, the averaged DSC over U-Net models that were trained on synthetic data from five different GANs for each ϵ is plotted. With decreasing ϵ , the DSC decreased and got more unstable, i.e., more variation between the different models for the same ϵ . In particular, models with $\epsilon > 5$

showed increased stability compared to models with lower ϵ . When considering only the best run of the five models (Figures 3B,C) the performance again dropped for decreasing ϵ . This was reflected by a lower DSC and a higher bAHD. The corresponding segmentation error maps are shown in Figure 4.

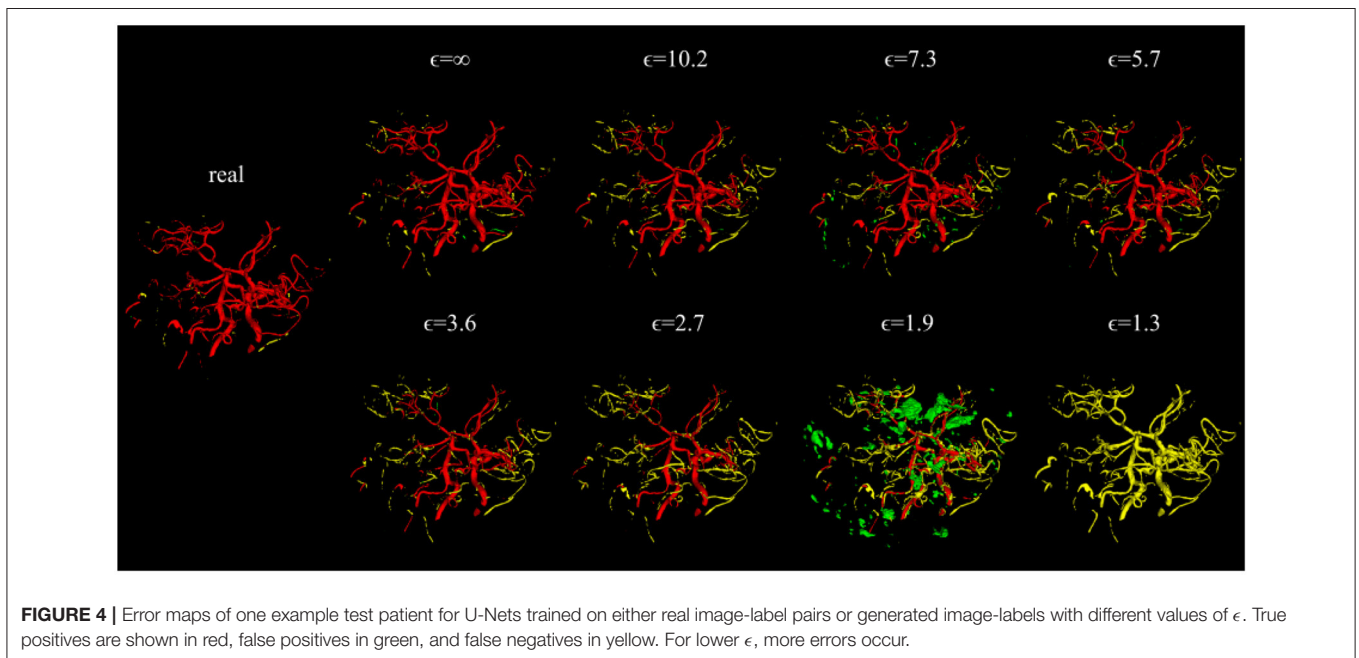
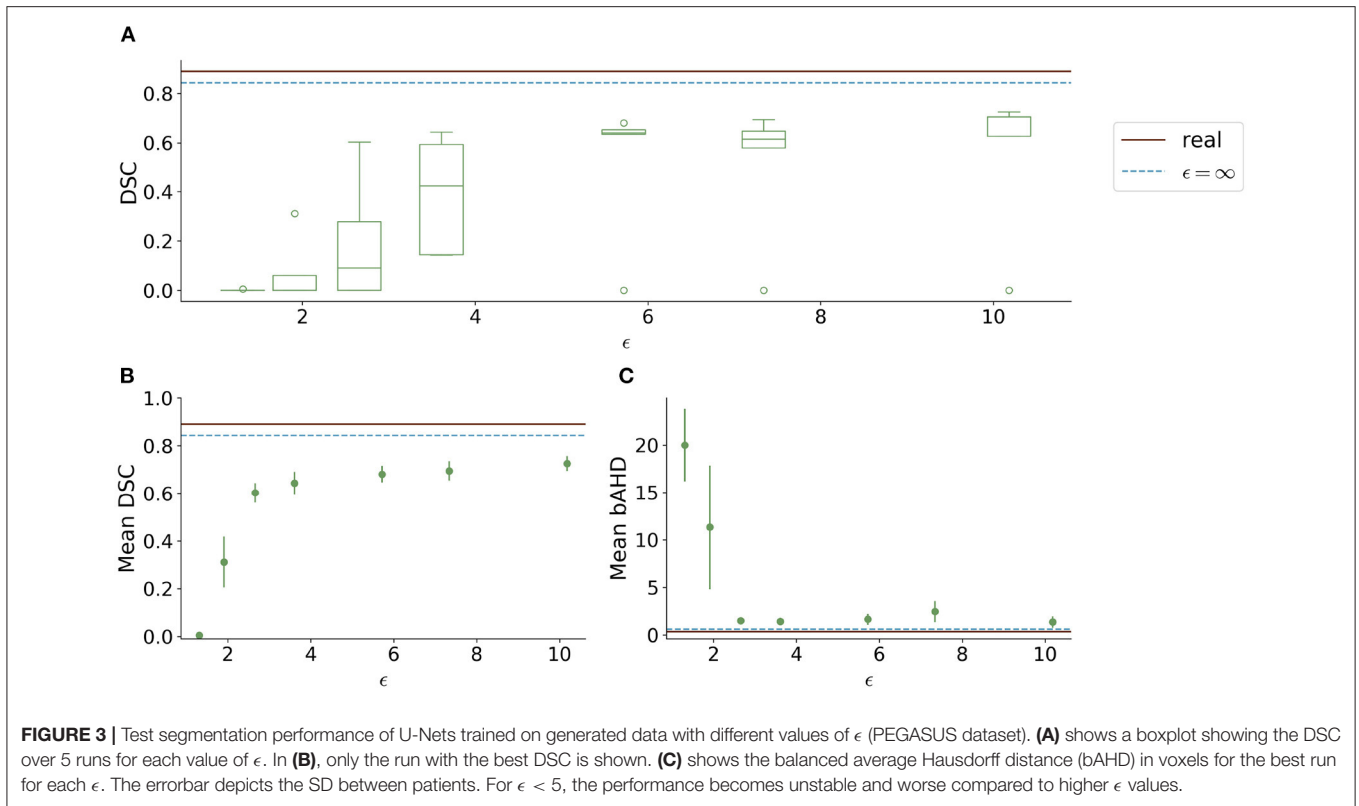
When testing the best U-Net models on the 1000Plus dataset, a similar trade-off between privacy and utility can be seen (Figure 5). Here, the U-Net performance in terms of DSC decreased more rapidly in comparison to the performance on the PEGASUS dataset, starting at $\epsilon = 8$ with DSC ≈ 0.69 (Figure 5A). The bAHD showed instability in performance for $\epsilon < 3$ (Figure 5B).

The FID between the training data and the generated data overall showed a similar trend: Less privacy led to a smaller distance to the training data (Figure 6A). The generated data trained without differential privacy ($\epsilon = \infty$) showed an FID of 62 compared to an FID of 244 and 228 for the images with $\epsilon = 5.7$ and $\epsilon = 10.2$, respectively. The distance to the test data was similar for different ϵ values. Figure 6B shows the difference between the distances to the training images and test images for different values of ϵ . Here, the differences were increasing for higher ϵ values with $\epsilon = \infty$ showing the largest difference, at least twice as large compared to all models trained with privacy guarantees.

Evaluating GAN models during training, we found the best performing image-label pairs when training with a noise multiplier of 0.65 for 29 epochs. This resulted in $\epsilon = 7.4$. The U-Net trained on these synthetic image-labels showed a DSC of 0.75 on the test set (Table 1). The segmentation of an example patient is shown in Figure 7. The big vessels are segmented reasonably well while a lot of errors occur when smaller vessels are segmented.

The similarity between the images is shown in Figure 8. For $\epsilon < 2$, high SSIM values were observed (SSIM > 0.98). In contrast, higher ϵ values led to less similar images produced by one model.

Figure 9 shows the t-SNE embedding of the best performing GAN with and without differential privacy and the real patient images. The synthetic images without privacy guarantees are overall close to the real images. The images with differential privacy cluster at the edges far away from the real images.



5. DISCUSSION

In the present study, we generated differentially private TOF-MRA images with corresponding labels and explored the trade-off between privacy and utility on two different test sets. We

proposed different evaluation schemes including training a segmentation network and identified a threshold of $\epsilon < 5$ with $DSC < 0.61$ for which the segmentation performance became unstable and not usable. Our best segmentation model trained on synthetic and private data achieved a DSC of 0.75 for

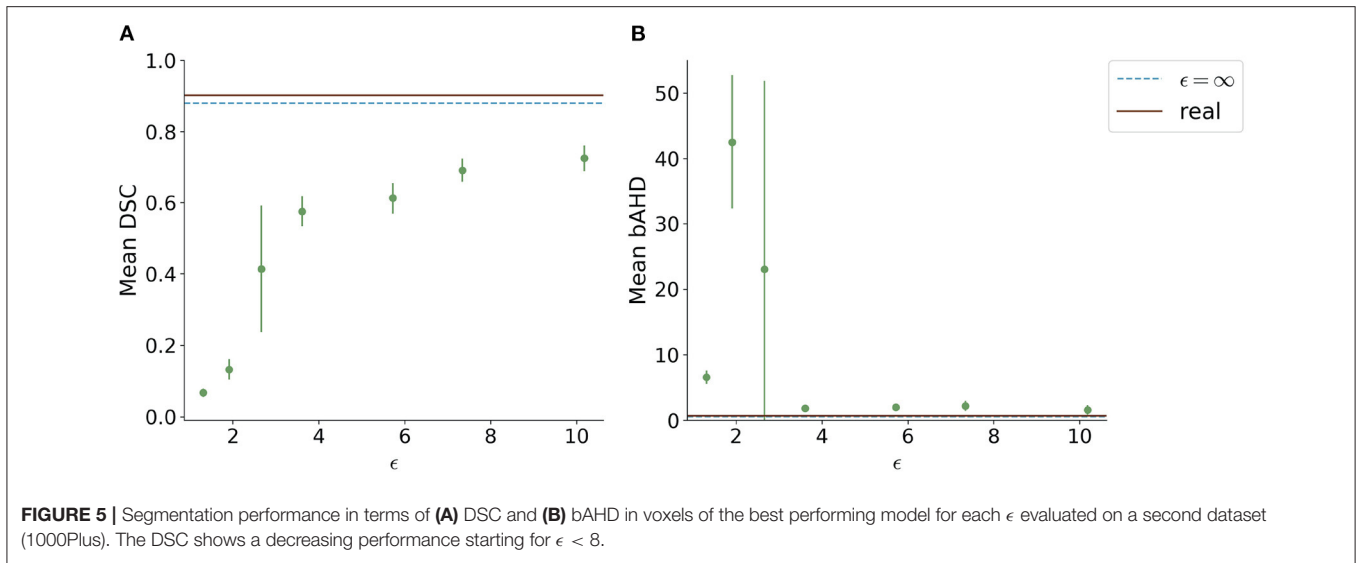


FIGURE 5 | Segmentation performance in terms of **(A)** DSC and **(B)** bAHD in voxels of the best performing model for each ϵ evaluated on a second dataset (1000Plus). The DSC shows a decreasing performance starting for $\epsilon < 8$.

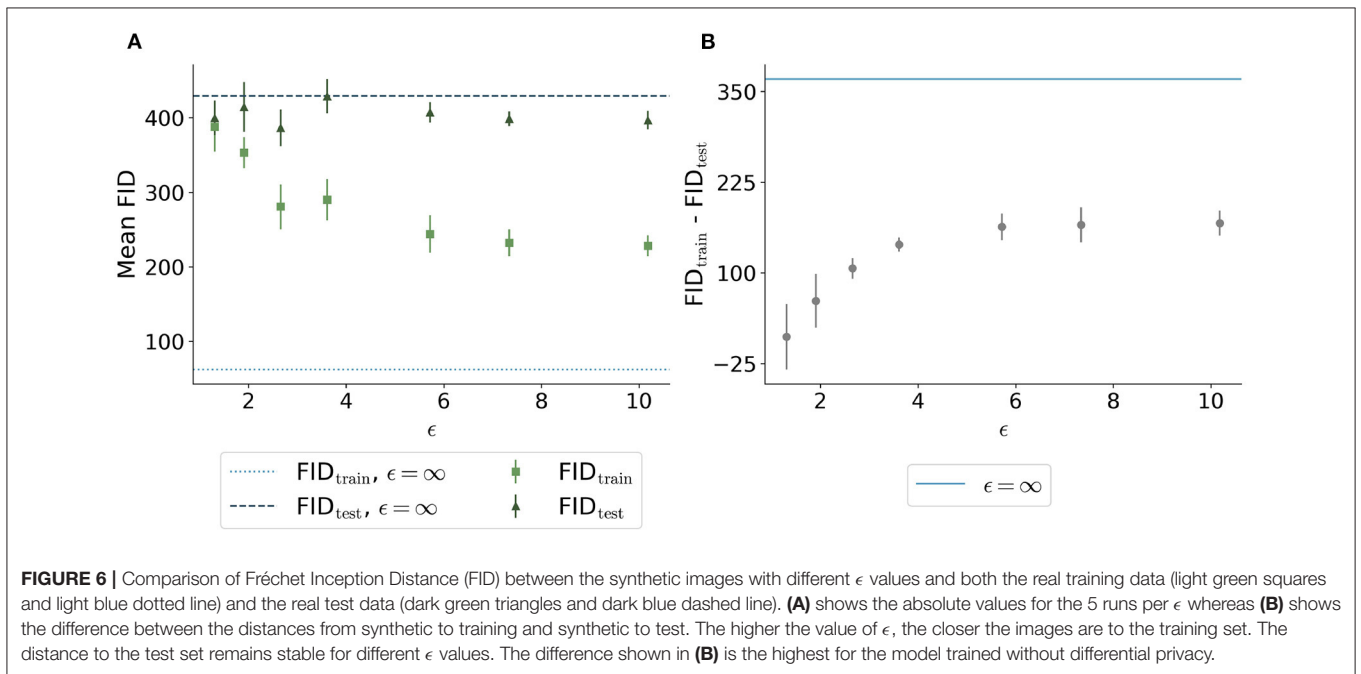


FIGURE 6 | Comparison of Fréchet Inception Distance (FID) between the synthetic images with different ϵ values and both the real training data (light green squares and light blue dotted line) and the real test data (dark green triangles and dark blue dashed line). **(A)** shows the absolute values for the 5 runs per ϵ whereas **(B)** shows the difference between the distances from synthetic to training and synthetic to test. The higher the value of ϵ , the closer the images are to the training set. The distance to the test set remains stable for different ϵ values. The difference shown in **(B)** is the highest for the model trained without differential privacy.

TABLE 1 | Overview of segmentation performances in terms of DSC and bAHD for a U-Net trained on real patient images and generated with and without differential privacy. The best of the three U-Net models is shown in bold for each metric and dataset. The best U-Net with differential privacy guarantees has an ϵ of 7.4. SD stands for standard deviation.

U-Net trained on	PEGASUS		1000Plus	
	Mean DSC (SD)	Mean bAHD (SD)	Mean DSC (SD)	Mean bAHD (SD)
Real images	0.89 (0.02)	0.33 (0.11)	0.90 (0.02)	0.69 (0.47)
Generated images ($\epsilon = \infty$)	0.84 (0.02)	0.61 (0.12)	0.88 (0.02)	0.58 (0.32)
Generated images ($\epsilon = 7.4$)	0.75 (0.04)	2.49 (1.96)	0.69 (0.04)	2.87 (1.25)

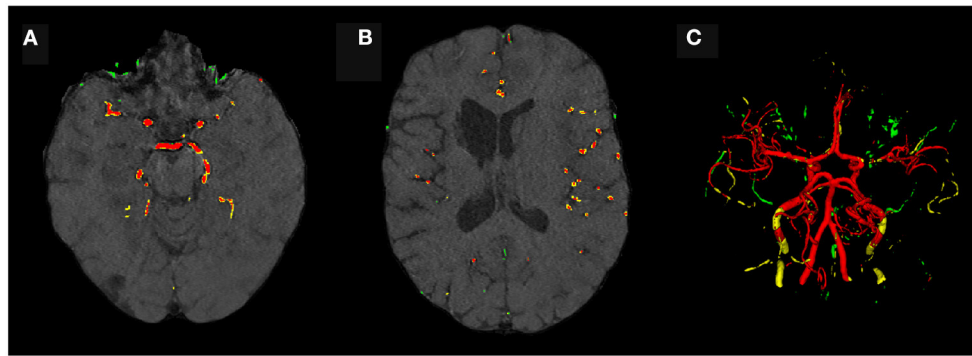


FIGURE 7 | Segmentation error maps of one test patient by the best U-Net model using differential privacy ($\epsilon = 7.4$). Red indicates the true positives, green stands for false positives, and yellow for false negatives. **(A)** shows a slice containing big vessels, **(B)** small ones, and **(C)** the whole vessel tree. The segmentation works reasonably well with errors occurring particularly when segmenting small vessels.

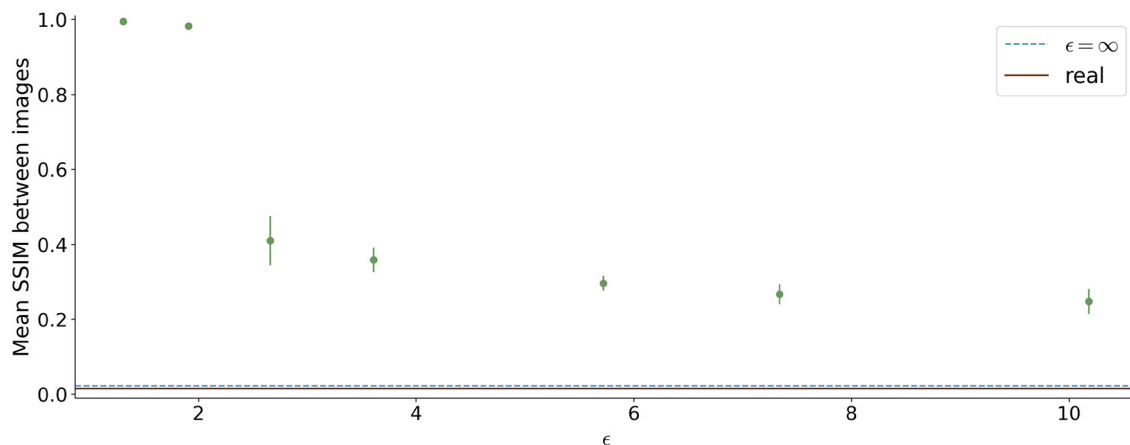


FIGURE 8 | Mean Structural Similarity Index Measure (SSIM) between 1,000 generated images for differential ϵ values. The errorbar shows the standard deviation over the 5 different runs for each ϵ value. For $\epsilon < 2$, the similarity between images is high, whereas it decreases for higher ϵ values.

$\epsilon = 7.4$ in a brain vessel segmentation paradigm. Our results mark the first step in data sharing with privacy guarantees for neuroimaging problems.

Since differential privacy is based on introducing noise, a decrease in utility is expected with the introduction of differential privacy. Our results confirm this notion. For $\epsilon = \infty$, we achieved a DSC of 0.84 which is comparable to the literature (Kossen et al., 2021). Stricter privacy constraints indicated by a lower ϵ led to worse visual results as well as poorer segmentation results (Figures 2–5). This also corresponds to findings in previous studies on differential privacy (Xie et al., 2018; Xu et al., 2019; Yoon et al., 2019). The increasing amount of noise might also be the reason for the instability of the GAN training for lower ϵ values, especially for $\epsilon < 5$ (Figure 2A). A performance drop could also be observed for testing the U-Nets trained on differential private image-label pairs on a second dataset (Figure 5). In comparison to the first test set, the performance drop occurred already for higher values of ϵ ($\epsilon < 8$ compared to $\epsilon < 5$). Thus, models with fewer privacy guarantees showed

better generalizability. A reason for that might be again the lower amount of noise and, therefore, fewer restrictions during training. This is also in line with our findings in Figure 8. Here, images generated from models with lower ϵ ($\epsilon < 2$) values showed more similarities between each other, thus indicating more mode collapse compared to models with higher ϵ values. This could be another reason for the performance drop for models with stricter privacy guarantees.

Images with larger ϵ values also showed greater similarity in terms of FID to the training images than those with stricter privacy guarantees. This indicates that more specific features of the training set can be memorized for less noisy models. The FID between test images and synthetic images (FID_{test}) stayed constant for different values of ϵ (Figure 6A). The difference between the FID_{train} and FID_{test} can be seen as a measure of the degree to which the images overfit the training set. Even for the model with our largest $\epsilon = 10.2$, the difference between FID_{train} and FID_{test} was only half compared to the difference of the model without any privacy constraints. This shows that

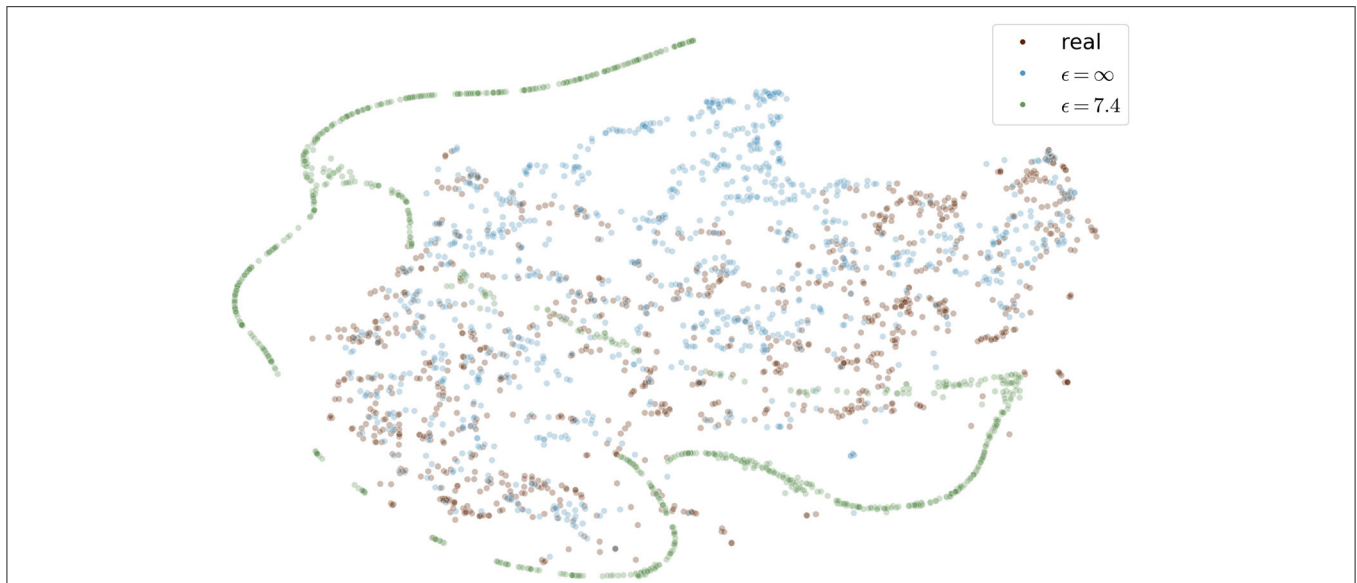


FIGURE 9 | Visualization of real and generated images with and without differential privacy in a t-SNE embedding. Each point represents an image. The distribution of real images and generated images without privacy almost entirely overlap. In contrast, the images with privacy guarantees are only partially overlapping and cluster at the edges, distant from the real images. The embedding showing the specific image instead of a point can be found in the **Figure S1** in the supplementary material.

differential privacy substantially contributed to the prevention of the memorization of the training set. Those findings are also in line with the embedding shown in **Figure 9** in which the differentially private images are further away from the training images compared to the images generated without any privacy guarantees.

Machine learning models including GANs are susceptible to so-called membership inference attacks (Shokri et al., 2017; Hayes et al., 2019; Chen et al., 2020). Here, an attack model is trained to predict whether a sample was part of the training set. If these attacks are successful, the privacy of the training samples is jeopardized. Differential privacy has been shown to decrease the model's vulnerability to privacy attacks (Shokri et al., 2017; Hayes et al., 2019). While there is no consensus about an exact value of ϵ , studies such as Hayes et al. (2019) and Bagdasaryan and Shmatikov (2019) consider a value of $\epsilon < 10$ acceptable. In this study, we were able to synthesize image-label pairs with single-digit ϵ (i.e., $\epsilon = 7.4$) that still show reasonable performance in the segmentation task. Naturally, further research is necessary to validate that our models would successfully defend against membership inference attacks.

Whereas, the segmentation performance in terms of DSC showed a consistent trend, this was not always true for the bAHD. **Figure 3C** shows overall comparable results to the DSC performance with some fluctuations. These fluctuations can be explained by selecting the best model based on the best validation DSC and not bAHD. In **Figure 5B**, however, the segmentation model for $\epsilon = 1.3$ seemed to perform better compared to models with $\epsilon = 1.9$ and $\epsilon = 2.7$. An explanation for this might be the number of false positives and false negatives in the segmentations. For $\epsilon = 1.3$, barely any voxel was identified as belonging to a vessel which resulted in many false

negatives. For the other two models, there were many false positives with a large distance to the ground truth. The bAHD considers these models to be worse although none of the three models show a good segmentation performance (see **Figure S2** in the supplementary material). The characteristic of penalizing especially false positives should be taken into consideration in future studies when using the bAVD as a metric.

The main limitations of the present study are the computational restrictions. Due to that only 2D patches were used. Additionally, more complex GAN architectures consisting of multiple generators and/or discriminators such as PrivGAN (Mukherjee et al., 2021) or PATE-GAN (Yoon et al., 2019) could not be implemented. Especially PrivGAN appears to be an interesting direction for future research since it does not only implement differential privacy but also aims to reduce vulnerability toward membership inference attacks directly.

6. CONCLUSION

In the present study, we synthesized differentially private TOF-MRA images and segmentation labels using GANs for a neuroimaging application. We proposed different evaluation metrics including the performance of a trained neural network for vessel segmentation. Even with privacy constraints, we could train a segmentation model that works reasonably well on real patient data. This is a crucial step toward synthesizing medical imaging data that both preserves predictive properties and privacy. Nonetheless, further studies should be conducted to evaluate if our findings generalize to other types of medical imaging data and to

further improve performance. Our synthetic data is available upon request.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The datasets used in this article are not readily available because data protection laws prohibit sharing the PEGASUS and 1000Plus datasets at the current time point. Requests to access these datasets should be directed to ethikkommission@charite.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Charité University Medicine Berlin and Berlin State Ethics Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TK, MH, VM, FB, KS, AHe, KH, SP, AHi, and DF: concept and design. VM, JS, IG, AK, and JF: acquisition of data. TK, VM, FB, AHe, KH, and DF: model design. TK: data analysis. TK, MH, VM,

FB, AHe, KH, and DF: data interpretation. TK, MH, VM, FB, KS, AHe, KH, SP, AHi, JS, IG, AK, JF, and DF: manuscript drafting and approval. All authors contributed to the article and approved the submitted version.

FUNDING

This study has received funding from the European Commission through a Horizon2020 grant (PRECISE4Q grant no. 777 107, coordinator: DF) and the German Federal Ministry of Education and Research through a Go-Bio grant (PREDICTioN2020 grant no. 031B0154 lead: DF).

ACKNOWLEDGMENTS

Computation has been performed on the HPC for the Research cluster of the Berlin Institute of Health. We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité-Universitätsmedizin Berlin.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2022.813842/full#supplementary-material>

REFERENCES

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., et al. (2016). "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16* (New York, NY: Association for Computing Machinery), 308–318.
- Abramian, D., and Eklund, A. (2019). "Refacing: reconstructing anonymized facial features using gans," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (Venice: IEEE).
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*. arXiv: 1701.07875.
- Aydin, O. U., Taha, A. A., Hilbert, A., Khalil, A. A., Galinovic, I., Fiebach, J. B., et al. (2021). On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking. *Eur. Radiol. Exp.* 5, 4. doi: 10.1186/s41747-020-00200-2
- Bagdasaryan, E., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. *CoRR, abs/1905.12101*.
- Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. (2019). Hypothesis testing interpretations and renyi differential privacy. *arXiv:1905.09982 [cs, stat]*. arXiv: 1905.09982.
- Bannier, E., Barker, G., Borghesani, V., Broeckx, N., Clement, P., Emblem, K. E., et al. (2021). The Open Brain Consent: Informing research participants and obtaining consent to share brain imaging data. *Hum. Brain Mapp.* 42, 1945–1951. doi: 10.1002/hbm.25351
- Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., et al. (2018). GAN Augmentation: augmenting training data using generative adversarial networks. *arXiv:1810.10863 [cs]*. arXiv: 1810.10863.
- Chen, D., Yu, N., Zhang, Y., and Fritz, M. (2020). "Gan-leaks: a taxonomy of membership inference attacks against generative models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, CCS '20* (New York, NY: Association for Computing Machinery), 343–362.
- Cheng, V., Suriyakumar, V. M., Dullerud, N., Joshi, S., and Ghassemi, M. (2021). "Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21* (New York, NY: Association for Computing Machinery), 149–160.
- Cirillo, M. D., Abramian, D., and Eklund, A. (2020). Vox2vox: 3d-gan for brain tumour segmentation. *CoRR, abs/2003.13653*. doi: 10.1007/978-3-030-72084-1_25
- Coyner, A. S., Chen, J. S., Chang, K., Singh, P., Ostmo, S., Chan, R. V. P., et al. (2022). Synthetic medical images for robust, privacy-preserving training of artificial intelligence: application to retinopathy of prematurity diagnosis. *Ophthalmol. Sci.* 2, 100126. doi: 10.1016/j.xops.2022.100126
- Duan, D., Xia, S., Reikik, I., Wu, Z., Wang, L., Lin, W., et al. (2020). Individual identification and individual variability analysis based on cortical folding features in developing infant singletons and twins. *Hum. Brain Mapp.* 41, 1985–2003. doi: 10.1002/hbm.24924
- Dwork, C. (2008). "Differential privacy: a survey of results," in *Theory and Applications of Models of Computation, Lecture Notes in Computer Science*, eds M. Agrawal, D. Du, Z. Duan, and A. Li (Berlin; Heidelberg: Springer), 1–19.
- Foroozandeh, M., and Eklund, A. (2020). Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE. *arXiv:2009.05946 [cs]*. arXiv: 2009.05946.
- Haarburger, C., Horst, N., Truhn, D., Broeckmann, M., Schradung, S., Kuhl, C., et al. (2019). "Multiparametric magnetic resonance image synthesis using generative adversarial networks," in *Eurographics Workshop on Visual Computing for Biology and Medicine* (The Eurographics Association Version Number: 011-015), 5.
- Hayes, J., Melis, L., Danezis, G., and Cristofaro, E. D. (2019). LOGAN: membership inference attacks against generative models. *Proc. Privacy Enhanc. Technol.* 2019, 133–152. doi: 10.2478/popets-2019-0008
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2018). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv:1706.08500 [cs, stat]*. arXiv: 1706.08500.
- Hilbert, A., Madai, V. I., Akay, E. M., Aydin, O. U., Behland, J., Sobesky, J., et al. (2020). Brave-net: Fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. *Front. Artif. Intell.* 3, 78. doi: 10.3389/frai.2020.552258

- Hotter, B., Pittl, S., Ebinger, M., Oepen, G., Jegzentis, K., Kudo, K., et al. (2009). Prospective study on the mismatch concept in acute stroke patients within the first 24 h after symptom onset-1000Plus study. *BMC Neurol.* 9, 60. doi: 10.1186/1471-2377-9-60
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2018). Image-to-image translation with conditional adversarial networks. *arXiv:1611.07004 [cs]*. doi: 10.1109/CVPR.2017.632
- Kossen, T., Subramaniam, P., Madai, V. I., Hennemuth, A., Hildebrand, K., Hilbert, A., et al. (2021). Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* 131, 104254. doi: 10.1016/j.combiomed.2021.104254
- Livne, M., Rieger, J., Aydin, O. U., Taha, A. A., Akay, E. M., Kossen, T., et al. (2019). A u-net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Front. Neurosci.* 13, 97. doi: 10.3389/fnins.2019.00097
- Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik* 29, 102–127. doi: 10.1016/j.zemedi.2018.11.002
- Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- Mironov, I. (2017). “Renyi differential privacy,” in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (Santa Barbara, CA: IEEE), 263–275.
- Mukherjee, S., Xu, Y., Trivedi, A., Patowary, N., and Ferres, J. L. (2021). privGAN: protecting GANs from membership inference attacks at low cost to utility. *Proc. Privacy Enhan. Technol.* 2021, 142–163. doi: 10.2478/popets-2021-0041
- Mutke, M. A., Madai, V. I., von Samson-Himmelstjerna, F. C., Zaro Weber, O., Revankar, G. S., Martin, S. Z., et al. (2014). Clinical evaluation of an arterial-spin-labeling product sequence in steno-occlusive disease of the brain. *PLoS ONE* 9, e87143. doi: 10.1371/journal.pone.0087143
- Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., and Zomaya, A. Y. (2021). Federated learning for COVID-19 detection with generative adversarial networks in edge cloud computing. *IEEE Internet Things J.* 1–1. doi: 10.1109/JIOT.2021.3120998
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *Mach. Learn. Python* 6, 2825–2830.
- Schwarz, C. G., Kremers, W. K., Therneau, T. M., Sharp, R. R., Gunter, J. L., Vemuri, P., et al. (2019). Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* 381, 1684–1686. doi: 10.1056/NEJMc1908881
- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). “Membership inference attacks against machine learning models,” in *2017 IEEE Symposium on Security and Privacy (SP)* (San Jose, CA: IEEE), 3–18.
- Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15, 29. doi: 10.1186/s12880-015-0068-x
- Torkzadehmahani, R., Kairouz, P., and Paten, B. (2019). “DP-CGAN: differentially private synthetic data and label generation,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Long Beach, CA: IEEE), 98–104.
- Tudosiu, P.-D., Varsavsky, T., Shaw, R., Graham, M., Nachev, P., Ourselin, S., et al. (2020). Neuromorphologically-preserving volumetric data encoding using VQ-VAE. *arXiv:2002.05692 [cs, eess, q-bio]*. arXiv: 2002.05692.
- Wang, L., Chen, W., Yang, W., Bi, F., and Yu, F. R. (2020). A State-of-the-Art review on image synthesis with generative adversarial networks. *IEEE Access* 8, 63514–63537. doi: 10.1109/ACCESS.2020.2982224
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861
- Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., et al. (2020). Preparing medical imaging data for machine learning. *Radiology* 295, 4–15. doi: 10.1148/radiol.2020192224
- Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network. *arXiv:1802.06739 [cs, stat]*. arXiv: 1802.06739.
- Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z., and Ren, K. (2019). GANobfuscator: mitigating information leakage under GAN via differential privacy. *IEEE Trans. Inf. Forensics Security* 14, 2358–2371. doi: 10.1109/TIFS.2019.2897874
- Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58, 101552. doi: 10.1016/j.media.2019.101552
- Yoon, J., Drumright, L. N., and van der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE J. Biomed. Health Inform.* 24, 2378–2388. doi: 10.1109/JBHI.2020.2980262
- Yoon, J., Jordon, J., and van der Schaar, M. (2019). “PATE-GAN: generating synthetic data with differential privacy guarantees,” in *International Conference on Learning Representations* (New Orleans: ICLR).
- Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., and Wu, Y. (2021). FedDPGAN: federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. *Inform. Syst. Front.* 23, 1403–1415. doi: 10.1007/s10796-021-10144-6
- Zhu, G., Jiang, B., Tong, L., Xie, Y., Zaharchuk, G., and Wintermark, M. (2019). Applications of deep learning to neuro-imaging techniques. *Front. Neurol.* 10, 869. doi: 10.3389/fneur.2019.00869

Conflict of Interest: TK, MH, VM, and AHi are employed by ai4medicine. FB and AHe are employed by Fraunhofer. JS reports receipt of speakers' honoraria from Pfizer, Boehringer Ingelheim, and Daiichi Sankyo. JF has received consulting and advisory board fees from BioClinica, Cerevast, Artemida, Brainomix, Biogen, BMS, EISAI, and Guerbet. DF receiving grants from the European Commission, reported receiving personal fees from and holding an equity interest in ai4medicine.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Kossen, Hirzel, Madai, Boenisch, Hennemuth, Hildebrand, Pokutta, Sharma, Hilbert, Sobesky, Galinovic, Khalil, Fiebach and Frey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

2.5 Comparing poor and favorable outcome prediction with machine learning after mechanical thrombectomy in acute ischemic stroke³⁶

Die zuvor beschriebenen Vorarbeiten flossen in die hier vorgestellte Arbeit ein, die erstmals die Grundlage für eine Machine Learning-basierte Entscheidungsunterstützung für die Behandlung des akuten Schlaganfalls legen konnte.

Vor dem Hintergrund, dass die Vorhersage der Ergebnisse nach mechanischer Thrombektomie (MT) bei Patienten mit akutem ischämischem Schlaganfall (AIS) und großem Gefäßverschluss (LVO) üblicherweise anhand eines guten Outcomes klassifiziert wird (modifizierte Rankin-Skala, mRS 0-2 nach 3 Monaten), stellt sich die Frage, ob diese Unterteilung für eine ML-basierte Unterstützung für die klinische Entscheidungsfindung geeignet ist.

Um diese Frage zu beantworten, analysierten wir retrospektiv Patienten mit akutem ischämischen Schlaganfall (AIS) und Okklusion eines großen Gefäßes (large vessel occlusion/LVO), die zwischen 2009 und 2018 eine mechanische Thrombektomie (MT) erhielten. Die prognostischen Variablen wurden in klinische Ausgangsvariablen, MRT-Variablen und Variablen, die die Geschwindigkeit und das Ausmaß der Reperfusion widerspiegeln (modified treatment in cerebral ischemia (mTICI) score), unterteilt.

Es wurden drei verschiedene Szenarien analysiert: (1) ausschließlich klinische Ausgangsparameter, (2) klinische und MRT-gestützte Ausgangsparameter und (3) alle klinischen, bildgebenden und reperfusionssassoziierten Ausgangsparameter. Für jedes Szenario wurde die Vorhersage von günstigen und schlechten Ergebnissen mit sieben verschiedenen maschinellen Lernalgorithmen bewertet.

Bei 210 Patienten verbesserte sich die Vorhersage des günstigen Ausgangs nach Einbeziehung der Geschwindigkeit und des Ausmaßes der Rekanalisation (AUC 0,73) im Vergleich zur ausschließlichen Verwendung der klinischen Variablen (AUC 0,67). Die Vorhersage eines schlechten Ergebnisses blieb bei ausschließlicher Verwendung der klinischen Ausgangsvariablen stabil (höchste AUC 0,71) und wurde durch die Addition zusätzlicher Variablen nicht weiter verbessert. Die Vorhersage der günstigen und schlechten Ergebnisse wurde durch die Hinzunahme von MR-Mismatch-Variablen nicht verbessert. Die wichtigsten klinischen Ausgangsvariablen für beide Ergebnisse waren Alter, National Institutes of Health Stroke Scale (NIHSS) und der modified Rankin Scale (mRS).

Unsere Ergebnisse deuten darauf hin, dass die Vorhersage eines schlechten Ergebnisses nach mit mechanischer Thrombektomie behandeltem Schlaganfall ausschließlich auf der Grundlage klinischer Ausgangsvariablen möglich ist. Die Schnelligkeit und der Umfang der Therapiemaßnahme verbesserten zwar die Vorhersage für ein günstiges Outcome, sind aber für ein schlechtes Outcome nicht relevant.



Comparing Poor and Favorable Outcome Prediction With Machine Learning After Mechanical Thrombectomy in Acute Ischemic Stroke

Matthias A. Mutke^{1*}, Vince I. Madaj^{2,3,4}, Adam Hilbert², Esra Zihni^{2,5}, Arne Potreck¹, Charlotte S. Weyland¹, Markus A. Möhlenbruch¹, Sabine Heiland¹, Peter A. Ringleb⁶, Simon Nagel⁶, Martin Bendszus¹ and Dietmar Frey²

¹ Department of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany, ² Charité Lab for Artificial Intelligence in Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany, ³ QUEST (Quality, Ethics, Open Science, Translation) Center for Responsible Research at Berlin Institute of Health, Charité Universitätsmedizin Berlin, Berlin, Germany, ⁴ School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, United Kingdom, ⁵ School of Computing, Technological University Dublin, Dublin, Ireland, ⁶ Department of Neurology, Heidelberg University Hospital, Heidelberg, Germany

OPEN ACCESS

Edited by:

Tae-Hee Cho,
Hospices Civils de Lyon, France

Reviewed by:

Bum Joon Kim,
University of Ulsan, South Korea
Carole Frindel,
Université de Lyon, France

*Correspondence:

Matthias A. Mutke
matthias.mutke@
med.uni-heidelberg.de

Specialty section:

This article was submitted to
Stroke,
a section of the journal
Frontiers in Neurology

Received: 07 July 2021

Accepted: 28 March 2022

Published: 27 May 2022

Citation:

Mutke MA, Madaj VI, Hilbert A, Zihni E, Potreck A, Weyland CS, Möhlenbruch MA, Heiland S, Ringleb PA, Nagel S, Bendszus M and Frey D (2022) Comparing Poor and Favorable Outcome Prediction With Machine Learning After Mechanical Thrombectomy in Acute Ischemic Stroke. *Front. Neurol.* 13:737667. doi: 10.3389/fneur.2022.737667

Background and Purpose: Outcome prediction after mechanical thrombectomy (MT) in patients with acute ischemic stroke (AIS) and large vessel occlusion (LVO) is commonly performed by focusing on favorable outcome (modified Rankin Scale, mRS 0–2) after 3 months but poor outcome representing severe disability and mortality (mRS 5 and 6) might be of equal importance for clinical decision-making.

Methods: We retrospectively analyzed patients with AIS and LVO undergoing MT from 2009 to 2018. Prognostic variables were grouped in baseline clinical (A), MRI-derived variables including mismatch [apparent diffusion coefficient (ADC) and time-to-maximum (Tmax) lesion volume] (B), and variables reflecting speed and extent of reperfusion (C) [modified treatment in cerebral ischemia (mTICI) score and time from onset to mTICI]. Three different scenarios were analyzed: (1) baseline clinical parameters only, (2) baseline clinical and MRI-derived parameters, and (3) all baseline clinical, imaging-derived, and reperfusion-associated parameters. For each scenario, we assessed prediction for favorable and poor outcome with seven different machine learning algorithms.

Results: In 210 patients, prediction of favorable outcome was improved after including speed and extent of recanalization [highest area under the curve (AUC) 0.73] compared to using baseline clinical variables only (highest AUC 0.67). Prediction of poor outcome remained stable by using baseline clinical variables only (highest AUC 0.71) and did not improve further by additional variables. Prediction of favorable and poor outcomes was not improved by adding MR-mismatch variables. Most important baseline clinical variables for both outcomes were age, National Institutes of Health Stroke Scale, and premorbid mRS.

Conclusions: Our results suggest that a prediction of poor outcome after AIS and MT could be made based on clinical baseline variables only. Speed and extent of MT did improve prediction for a favorable outcome but is not relevant for poor outcome. An MR mismatch with small ischemic core and larger penumbral tissue showed no predictive importance.

Keywords: stroke, mechanical thrombectomy, outcome prediction, machine learning, MRI, perfusion imaging, mismatch

INTRODUCTION

Mechanical thrombectomy (MT) is the most effective treatment for patients with acute ischemic stroke (AIS) due to a large vessel occlusion of the anterior circulation (1). While the average treatment effect and outcome benefit across the entire group of patients is large, outcome still differs significantly for individual patients (1). Multiple prognostic variables and their combination render individual outcome prognosis after MT difficult. For example, in a group of patients with successful and fast reperfusion, ~60% still had an unfavorable outcome (mRS 3–6 after 3 months) (2). At present, the relative importance and combination of single prognostic variables for individual outcome prediction is still a matter of debate.

One way to address this problem is to utilize artificial intelligence, in particular machine learning (ML) approaches. These models are potentially superior to conventional linear or logistic regression models as they excel at finding complex and non-linear relationships across a multitude of prognostic variables. Specially, artificial neural networks and methods of tree-boosting are promising tools in this regard (3). Recent advances have made it possible to uncover which individual prognostic variables are most important in such models (4, 5) based on a feature importance analysis.

Applying this methodology to outcome prediction after MT, multiple prognostic variables can be used representing the clinical course of stroke patients: baseline clinical variables (1), MRI variables including perfusion and infarct core (6), and finally, variables assessing the speed and extent of reperfusion (7).

The outcome and potential benefit of MT are usually assessed after 3 months with the modified Rankin Scale (mRS) in a dichotomized analysis: 0–2 is defined as favorable outcome and the remaining 3–6 as unfavorable outcome. Patients with a predicted favorable outcome will undoubtedly undergo MT. However, the remaining group of patients with unfavorable outcome is highly heterogeneous, ranging from outcomes of mRS of 3 (moderate disability) to 6 (death).

Therefore, it may also be reasonable to find prognostic factors for a poor outcome (8) (severe disability or death after 3 months, with an mRS score of 5 or 6). In those patients, withholding treatment could be discussed.

Therefore, in the presented work we used ML to predict outcome after MT directly comparing two different dichotomization paradigms: for favorable (mRS 0–2 vs. 3–6) and poor outcome (mRS 5 and 6 vs. 1–4) with multiple prognostic variables grouped in three sets: baseline clinical, MRI-derived, and MT-associated variables.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

The study protocol for this retrospective analysis of our prospectively established stroke database was approved by the ethics committee of Heidelberg University and patient-informed consent was waived.

Patients

We identified patients with AIS due to an occlusion of the middle cerebral artery in the M1 or M2 segment or the distal terminus of the internal carotid artery who were treated with MT between 03/2009 and 09/2018; 95/210 patients were treated between 2010 and 2013, and the remaining 116/210 between 2014 and 2018. Between these two groups, there was no significant outcome difference after 3 months (Mann–Whitney test, $p = 0.83$).

Patients were treated at a single center (University Hospital Heidelberg). The attending neurologist and interventional neuroradiologist decided on treatment on a case-by-case basis. Only patients with a completed MRI protocol and outcome assessment at 3 months were included.

Baseline clinical and imaging parameters are given in **Table 1**. Individual patient outcome was the score on the mRS (9) at 3 months assessed by a standardized interview (unblinded investigator per phone call or a personal letter to the patient). The mTICI was used to grade recanalization on final angiographic images (10). A score of mTICI 2b or better on final angiogram was regarded as successful reperfusion.

MRI Protocol

In a routine clinical setting, MR images were acquired on 3 Tesla MRI systems (Magnetom Verio, TIM Trio and Magnetom Prisma; Siemens Healthcare, Erlangen, Germany). Imaging protocol included diffusion-weighted, FLAIR, susceptibility

Abbreviations: MT, mechanical thrombectomy; AIS, acute ischemic stroke; LVO, large vessel occlusion; mRS, modified Rankin Scale; DWI, diffusion weighted imaging; MRI, magnetic resonance imaging; Tmax, time-to-maximum; mTICI, modified treatment in cerebral ischemia; NIHSS, National Institutes of Health Stroke Scale; GLM, generalized linear model; SVMC, Support Vector Machine Classifier; NB, Naive Bayes; MLP, Multilayer Perceptron; AUC, area-under-the-curve; ROC, receiver-operating-characteristic; SHAP, Shapley Additive Explanations; VIF, variance inflation factor.

TABLE 1 | Prognostic variables (features).

		All patients (n = 210)	Favorable outcome (n = 83/210)	Poor outcome (n = 49/210)
Set A: Baseline clinical variables				
Time from stroke onset to MR-imaging (time in minutes)	Median/IQR	260 (126–569)	257 (123–552)	219 (109–526)
Wake up stroke	%	85 (40%)	31 (37%)	21 (43%)
Age (years)	Median/IQR	72 (59–78)	69 (53–75)	76 (68–81)
Sex	Male/female	92/118	39/44	21/28
Diabetes	%	37 (18%)	7 (8%)	14 (29%)
Hypertonia	%	131 (62%)	47 (57%)	37 (76%)
Coronary heart disease	%	36 (17%)	11 (13%)	12 (24%)
Arrhythmia/atrial fibrillation	%	77 (37%)	22 (27%)	24 (49%)
Hyperlipidemia	%	62 (30%)	23 (28%)	16 (33%)
NIHSS scale at admission (0–42)	Median/IQR	16 (12–20)	15 (10–20)	20 (15–30)
mRS pre-stroke (0–5)	Median/IQR	0 (0–1)	0 (0–1)	1 (0–2)
i.v. Thrombolysis	%	132 (63%)	58 (70%)	29 (59%)
Set B: Magnetic resonance imaging derived variables				
ADC lesion volume (ml)	Median/IQR	14 (8–30)	16 (7–32)	14 (9–27)
Tmax lesion volume (ml)	Median/IQR	78 (39–140)	73 (36–121)	105 (60–168)
Mismatch ratio (Tmax lesion volume/ADC lesion volume)	Median/IQR	4.6 (2.3–8.4)	4.2 (2.1–8.4)	5.3 (2.9–13.3)
Occlusion distal carotid artery	%	12 (6%)	7 (8%)	2 (4%)
Occlusion carotid terminus	%	33 (16%)	14 (17%)	7 (14%)
Occlusion M1 segment middle cerebral artery	%	131 (62%)	49 (59%)	32 (65%)
Occlusion M2 segment middle cerebral artery	%	36 (17%)	14 (17%)	7 (14%)
Set C: Thrombectomy associated variables				
Final mTICI score (TICI 3 and 2b)	%	154 (73%)	73 (88%)	26 (53%)
Time from stroke onset to final mTICI score	Median/IQR	492 (330–787)	526 (319–853)	434 (319–721)

Prognostic variables were grouped in three distinct sets: Baseline clinical variables (A), MRI-derived mismatch variables (B), and mechanical thrombectomy-associated variables (C). Variables are given for all patients and the two subgroups of patients with favorable outcome (mRS 0–2) and with poor outcome (mRS 5 and 6) after 3 months.

weighted and T2-weighted sequences, non-contrast time-of-flight, and contrast-enhanced angiography as well as dynamic susceptibility contrast perfusion-weighted imaging. The imaging protocol has been published previously and is included in the **Supplementary Material** (11).

Image Post-processing

All image analysis was performed blinded to clinical outcome. Diffusion-weighted imaging (DWI) and perfusion MRI images were post-processed with Olea Sphere[®] (Olea Medical[®], La Ciotat, France). ADC maps were automatically calculated from DWI images with different *b*-values. For perfusion imaging, automatic motion correction was applied. The arterial input function (AIF) was detected automatically. In two cases, the automatically detected AIF was manually corrected. Tmax maps were calculated using a block-circulant singular-value decomposition (cSVD) deconvolution algorithm. Diffusion lesion volumes [ADC value threshold of $\leq 620 \times 10^{-6} \text{ mm}^2/\text{s}$ (12)] and Tmax lesion volumes [Tmax threshold $\geq 6 \text{ s}$ (13)] were segmented semiautomatically and manually corrected for artifacts by a neuroradiologist (MM) with more than 6 years of experience in stroke imaging.

ML Framework

For the training of the ML models, we utilized a publicly available ML framework for predictive modeling. The program code is available on Github (<https://github.com/prediction2020/explainable-predictive-models>). Details on the technical implementation have been described in open-access publications previously (5).

Definition of Prognostic Paradigms

We defined two distinct prognostic paradigms: For the first paradigm **I**, all patients included in the study were dichotomized in favorable outcome with an mRS of 0, 1, or 2 at 3 months vs. the remaining with mRS 3–6. For the second paradigm **II**, again all patients were included and dichotomized but in poor outcome, defined as mRS 5 or 6 at 3 months vs. the remaining with mRS 0–4. The dichotomized mRS was used as a label for the ML analysis.

Prognostic Variables for Input Feature Definition

We grouped prognostic variables in three distinct sets (**Table 1**): Baseline clinical variables (A), MRI-derived variables (B), and

thrombectomy-related variables (C). The mTICI score as a measure for the success of reperfusion was dichotomized for the final analysis (mTICI 2b/3 vs. 0–2a). The prognostic variables included in the three sets were used as input features for the ML analysis. Mismatch ratio was not included as an independent feature because it is derived from the ADC and TMax lesion volumes and would be redundant information. Target mismatch was defined according to the EXTEND-IA study (6) with an infarct core of <70 ml on ADC maps, a ratio of TMax Lesion volume to ADC lesion volume of 1.2 or higher and an absolute mismatch volume of 10 ml or more.

We defined three distinct prediction scenarios with the different sets of prognostic variables: Prediction with baseline clinical variables only (A), with baseline clinical and MRI variables combined (A+B), and finally with all baseline clinical, MRI and thrombectomy-associated variables combined (A+B+C).

For each of the three scenarios, we trained models for both prognostic dichotomization paradigms I and II with favorable (mRS ≤ 2) and poor outcome (mRS 5 or death), respectively. This yielded six different scenarios in total (see **Figure 1**).

Multicollinearity was estimated using the variance inflation factor (VIF) (5).

Applied Algorithms

We utilized all seven available ML algorithms from the framework to provide a comprehensive coverage of various ML methods. The more traditional techniques were represented by three algorithms: A generalized linear model (GLM), which for dichotomous outcomes is equivalent to a plain logistic regression, and two regularized variants, a Lasso algorithm with L1 regularization and an ElasticNet with L1 and L2 regularization.

Further, ML algorithms included tree boosting (Catboost implementation), a Support Vector Machine Classifier (SVMC), Naive Bayes (NB), and a Multilayer Perceptron (MLP).

Model Training and Validation

The data comprising the given clinical parameters and outcomes were randomly split into training and test sets in a corresponding 4:1 ratio. Due to slight imbalance with respect to the outcome measures (127/210 patients with favorable outcome in paradigm I and 49/210 patients with poor outcome in paradigm II), random sub-sampling of the majority class was employed for the training sets. Test sets were stratified to follow the original imbalanced ratio to represent real distribution of our patient outcomes in model testing. In total, there were only 11 missing values in the data set. Missing values were imputed by mean/mode imputation. Non-categorical features—both in training and test sets—were standardized to zero-mean and unit variance based on training set statistics. Models were trained and best parameters were selected using 10-fold cross validation over an extensive grid of hyperparameters for each model. Parameter ranges were initially taken from the public repository referenced under the heading “ML framework”, and then refined taking run times of experiments into consideration. The used ranges are included in **Supplementary Material**. The whole process was repeated

200 times (shuffles) to account for dependence on the random procedure of train and test splits.

Performance Assessment

For performance measures, we report results as the median over the test sets of the 200 shuffles. Model performance was primarily assessed by area-under-the-curve (AUC) measure resulting from receiver-operating-characteristic (ROC) analysis. Accuracy, balanced class accuracy, precision, recall, f1 score, negative predictive value, and specificity measures for each model are included in the **Supplementary Material**. Statistical significance of the difference between model performances on the respective variable sets was determined by the Wilcoxon signed-rank test at a confidence level of 5%.

Explainability Assessment

We used SHapley Additive exPlanations (SHAP) scores to rate the importance of included clinical features for all seven models. More detailed explanation of the technique can be found in (14). The absolute values of importance scores on test sets were scaled to unit norm to yield comparable measures for all models, and then rescaled to the range of [0, 1] so that importance scores for a certain model sum to 1. Finally, mean and standard deviation across the 200 shuffles for stability and robustness were calculated and reported as the final importance rating.

RESULTS

Patients

In total, 236 patients met the inclusion criteria, and 26 patients were excluded due to motion artifacts on MRI images or because no accurate AIF could be obtained, resulting in the final number of 210 patients. Median mismatch ratio was 4.6 (2.3–8.4). In 154 patients (73%), successful reperfusion (TICI 3 or 2b) could be achieved. Median time to TICI was 492 min. In prediction paradigm I, 83/210 patients (39%) had a favorable outcome (mRS 0–2). In prediction paradigm II, 49/210 patients (23%) had a poor outcome (mRS 5–6).

In this study, 168/210 patients (80%) had a target MRI mismatch [according to the EXTEND-IA study criteria (6)]; 5/210 patients (2%) had no ischemic core and 11/210 patients (5%) without target mismatch had a small ischemic core of 10 ml or less.

In the multicollinearity analysis, VIF values were below 5 for all scenarios using the predictive variable sets A and A+B. For the two scenarios I A+B+C and II A+B+C, time from stroke onset to final TICI and time from onset to MRI raised to values ~ 9.9 , indicating stronger multicollinearity for these features. We did not recognize a harmful level of multicollinearity in any of the variable sets; thus, no features were eliminated.

Prediction Models

The specific AUC results for the total of six prediction scenarios, each examined with seven algorithms are presented

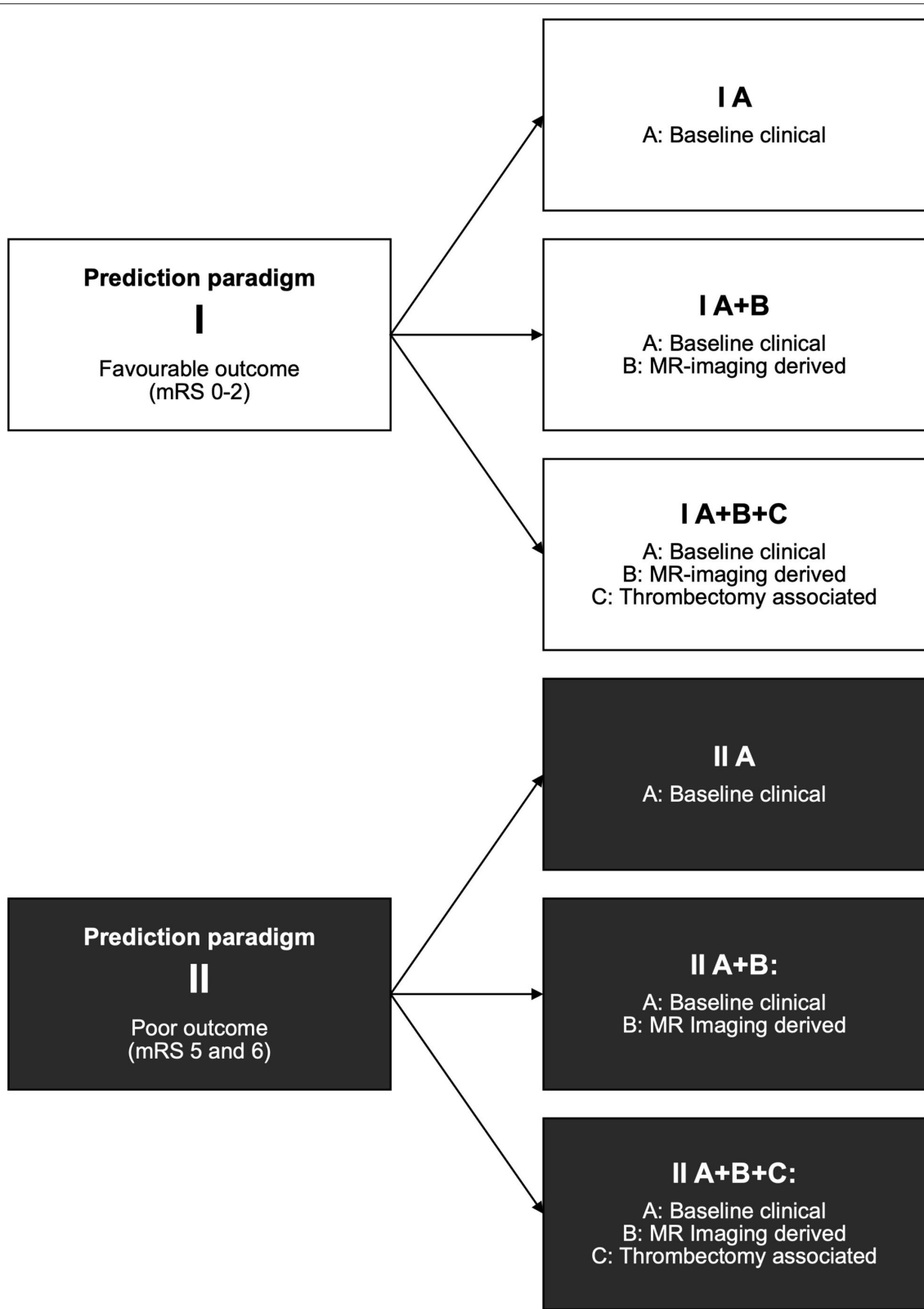


FIGURE 1 | Prediction paradigms and resulting scenarios. For each paradigm, all patients included in the study were dichotomized: Paradigm I for favorable outcome with mRS 0–2 at 3 months (vs. the remaining 3–6) and paradigm II for poor outcome with mRS 5 and 6 (vs. the remaining 0–4). For the prediction scenarios, three sets of prediction variables A, B, and C were consecutively added. For an overview of prediction variables included in the sets, see **Table 1**. The combination of each of the three prediction variable sets and two prediction paradigms yielded six distinct scenarios.

TABLE 2 | Models for favorable outcome (paradigm I).

Scenario	GLM	Lasso	ElasticNet	Catboost	MLP	SVMC	Naive bayes
I A	0.65	0.65	0.6	0.67	0.67	0.6	0.65
I A+B	0.62*	0.64*	0.6	0.64*	0.64*	0.57	0.63*
I A+B+C	0.71*+	0.71*+	0.68*+	0.73*+	0.7*+	0.67*+	0.69*+

The prediction variable sets (Table 1) were used to predict a favorable outcome (mRS dichotomized as 0–2 vs. 3–6). The addition of thrombectomy-associated variables (set C) leads to a noticeable improvement of all machine learning models. Highest AUC results for each variable set are marked in bold. Confidence intervals for all models are included in the **Supplementary Material**. Statistically significant difference in model performance between variable sets are marked with * and + to signal difference from A and from A+B, respectively. Significance was determined by a value of *p* lower than 0.05, resulting from the Wilcoxon signed-rank test.

TABLE 3 | Models for poor outcome (paradigm II).

Scenario	GLM	Lasso	ElasticNet	Catboost	MLP	SVMC	Naive bayes
II A	0.67	0.7	0.64	0.7	0.71	0.59	0.69
II A+B	0.65*	0.7	0.62*	0.7	0.69	0.57	0.65*
II A+B+C	0.68+	0.71	0.65+	0.73*+	0.7	0.65*+	0.66*

The prediction variable sets (Table 1) were used to predict poor outcome (mRS dichotomized as 5 and 6 vs. 0–4). In contrast to the favorable outcome paradigm, the addition of thrombectomy-associated variables (set C) did not lead to relevant improvements in the performance of machine learning models. Only the Catboost model profited slightly in a clinically relevant AUC range. Highest AUC results for each variable set are marked in bold. Confidence intervals for all models are included in the **Supplementary Material**. Statistically significant difference in model performance between variable sets are marked with * and + to signal difference from A and from A+B, respectively. Significance was determined by a value of *p* lower than 0.05, resulting from the Wilcoxon signed-rank test.

in Table 2 for paradigm I with favorable outcome and in Table 3 for paradigm II with poor outcome, respectively. The results for the additional performance measures are given in **Supplementary Material**.

For the first scenario with baseline clinical variables only (I A and II A), prediction was slightly better for poor outcome (II A) than for favorable outcome (I A). The smallest difference in AUC between I A and II A was 0.02 for GLM and the largest 0.05 for Lasso logistic regression. Only SVMC showed comparable results.

Adding MRI-derived parameters (scenario I A+B and II A+B) did not change the prediction performance for both paradigms. This was consistent for all algorithms in both scenarios I A+B and II A+B.

Finally, adding thrombectomy-associated parameters—extent and speed of recanalization—I A+B+C and II A+B+C improved the prediction performance noticeably across many algorithms for the favorable outcome paradigm (I A+B+C). Prediction for the poor outcome paradigm with all variables (II A+B+C) remained approximately stable; only the Catboost and SVMC algorithm showed a slight improvement (AUC increase of 0.03, 0.08, respectively, compared to II A).

To summarize, prediction for the poor outcome paradigm II remained comparable on a relatively high level across all three prediction scenarios (II A, II A+B, II A+B+C). Contrariwise, prediction for the favorable outcome paradigm I improved noticeably when thrombectomy-associated parameters were added (I A+B+C). The final performance for the last scenario with all predictive variables included (I A+B+C and II A+B+C) was comparable for both the favorable and poor paradigms.

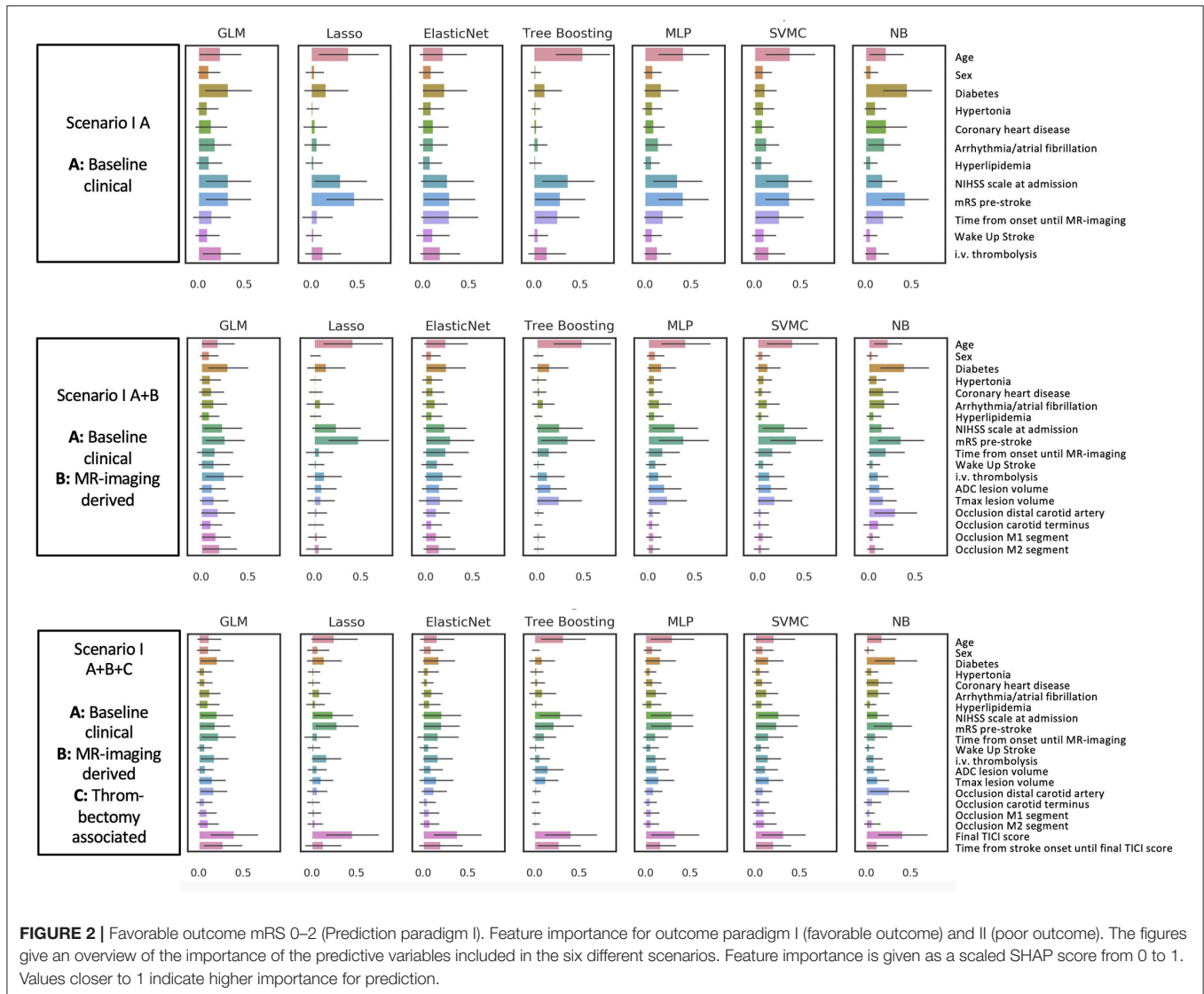
Feature Importance Ranking

Feature importance values for each scenario and each algorithm are displayed in **Figures 2, 3**.

For the favorable outcome paradigm I (Figure 2), the most relevant variables across all algorithms for the first scenario with the baseline clinical variables only (I A) were age, pre-stroke mRS, National Institutes of Health Stroke Scale (NIHSS) at admission, and time from stroke onset to MRI. Adding MRI-derived mismatch parameters and the site of occlusion (I A+B), the Tmax volume for hypoperfused tissue was of higher relevance, while the ADC volume was of moderate importance. However, this was only visible in some models, among them the tree boosting model (the best performing model in the favorable outcome paradigm I). For models with all variables combined (I A+B+C), the mTICI score became the most dominant parameter in all models. Also, time from stroke onset to final TICI score was assigned high importance by the majority of models.

For the poor outcome prediction paradigm (Figure 3), we found a similar pattern. Age, premorbid mRS, and the baseline NIHSS were the most relevant features in the model with baseline variables only (II A). However, i.v. thrombolysis and risk factors such as diabetes played a smaller role compared to the favorable outcome paradigm (I A). For the models with additional MR-mismatch parameters (II A+B), both ADC and Tmax volume were of less importance than in the favorable outcome paradigm (I A+B). In the third scenario with all variables included (II A+B+C), the baseline variables from the first scenario (II A) remained important. Additionally, the mTICI score was relevant in most models, however not as relevant as compared to the favorable outcome prediction (I A+B+C).

To summarize, for the prediction of either good or poor outcome, age, premorbid mRS, and baseline NIHSS were



important, with mTICI score as an additional relevant feature from the third scenario. ADC and Tmax volume were more important for the favorable than for the poor outcome paradigm. Information about i.v. thrombolysis was only important for the favorable outcome paradigm.

DISCUSSION

In this study, we examined ML-based outcome prediction models for patients with stroke who underwent MT. We compared prediction of poor outcome (mRS 5 or 6 vs. 0–4) and favorable outcome (mRS 0–2 vs. 3–6) measured at 3 months post stroke. These prediction paradigms have direct implications for clinical decision-making by predicting an outcome of no or only slight disability on the one hand and severe disability or death on the other. In particular, the definition of favorable outcome is generally accepted and was applied in large prospective studies.

We chose different combinations of prognostic variables that were deliberately limited to those most commonly used in stroke practice and literature (1, 6, 7) and most accessible in clinical decision-making, especially under time constraints as encountered in clinical practice. We found considerable differences between the two outcome paradigms.

Our main finding suggests that prediction of poor outcome may possibly be based on clinical baseline variables only and set a rather high benchmark in the first prediction scenario. The predictive performance did not improve by adding target MR-mismatch and recanalization-related parameters. In contrast, prediction of favorable outcome did improve significantly by adding speed and extent of recanalization compared to using baseline clinical variables only.

In contrast to previous studies, the main strength of the presented work is the direct comparison of two different outcome prediction paradigms. The choice of how to dichotomize mRS for outcome prediction has clinical relevance: The standard

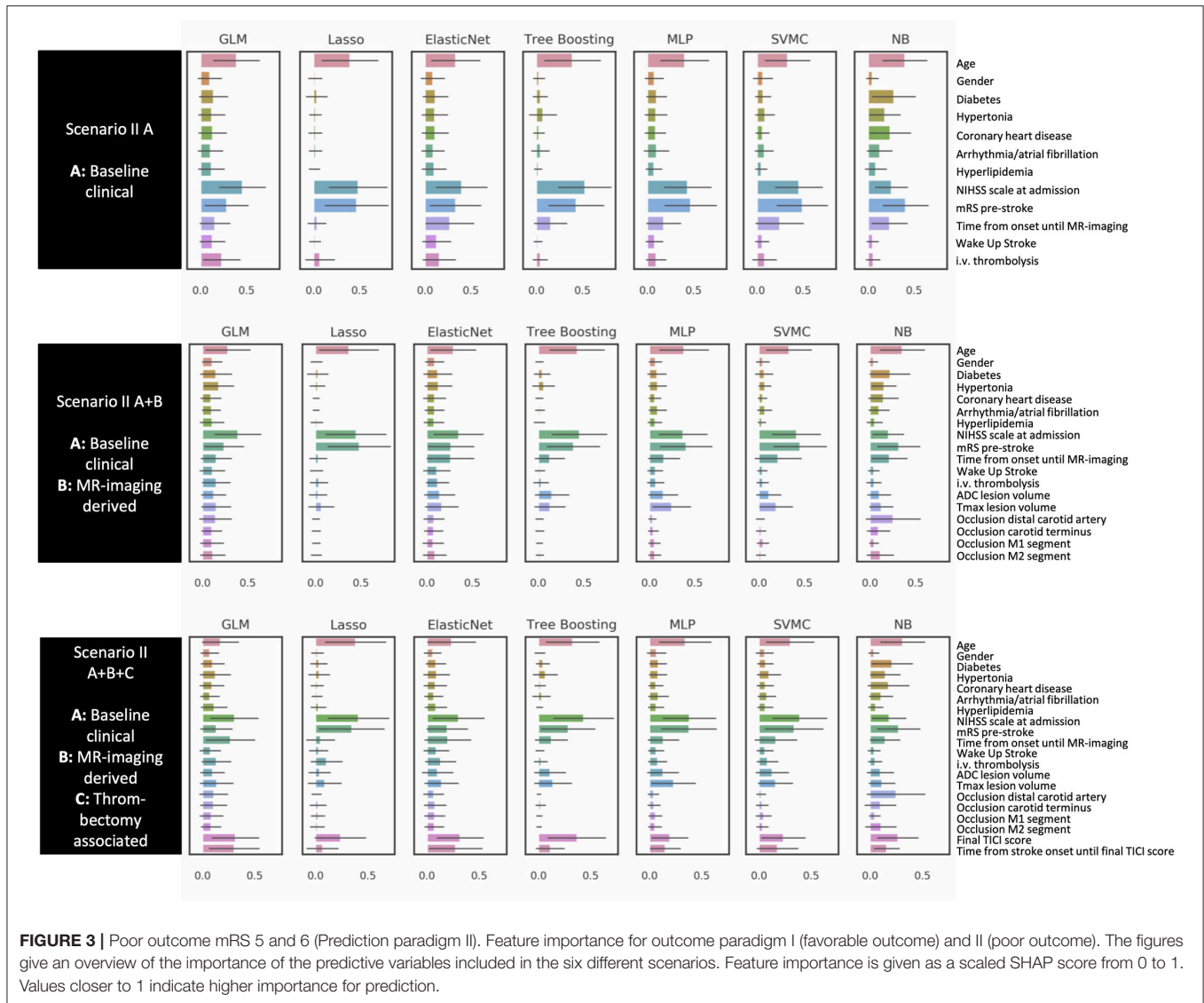


FIGURE 3 | Poor outcome mRS 5 and 6 (Prediction paradigm II). Feature importance for outcome paradigm I (favorable outcome) and II (poor outcome). The figures give an overview of the importance of the predictive variables included in the six different scenarios. Feature importance is given as a scaled SHAP score from 0 to 1. Values closer to 1 indicate higher importance for prediction.

dichotomization with 0–2 vs. 3–6 as used in the large randomized clinical trials puts an emphasis on patients with favorable outcomes. Dichotomizing mRS 0–4 vs. 5–6 focuses on patients with a very high degree of dependency or mortality after MT.

Our exploratory analysis revealed that patients with expected poor outcome could already be captured by clinical baseline variables before thrombectomy. The prediction scenario with baseline clinical variables only was already close to the final performance when mismatch and recanalization information was added. In those patients, withholding MT based on the clinical baseline variables could be the consequence; however, these implications should be verified in larger, prospective studies (8) and within new clinical data sets.

This was in contrast to the prediction of favorable outcome, where the prediction with baseline clinical variables only

was lower but could be considerably improved by adding information about the speed and extent of reperfusion. However, because of the retrospective and thus exploratory nature of our study, our results should be interpreted with caution with regards to clinical treatment decisions. Nonetheless, they warrant further exploration in prospective studies to confirm our findings. Within such prospective data, the proposed paradigms could then be used to estimate individualized chances for either poor or favorable outcome before and after therapy. A similar approach using two models for an individualized prognosis of the same outcome variable is proposed by Debs et al. (15).

This performance pattern was accompanied by complimentary information from the feature importance analysis: For poor outcome, age, stroke severity (NIHSS) and degree of disability before stroke (premorbid mRS before stroke), and the time from stroke onset to imaging were the most

important baseline variables. For favorable outcome, the most important predictive variables were also age, NIHSS at onset, and the mRS before stroke, but additionally, the speed and extent of recanalization (mTICI score and the time from stroke onset to the final mTICI score) were paramount. NIHSS was a more important predictor for final outcome than the ischemic lesion volume before therapy. This could be due to lesion location in eloquent brain regions where a smaller infarct causes comparably more severe clinical symptoms. Also, the final ischemic lesion volume after therapy might improve prediction but was not available to be included in our models.

Across all models, the maximum predictive value was an AUC of 0.73 for both favorable and poor outcome with regards to model performance and feature rankings. Our findings are comparable to previous studies applying ML algorithms: Hammam et al. (16) found a similar prediction for favorable outcome for patients with MT which did not considerably improve by adding MR-mismatch and other imaging-derived parameters. Other studies applied ML for outcome prediction with baseline CT imaging: Brugnara et al. (17) found an AUC of 0.85 for the prediction of favorable outcome only after adding information about infarct size after thrombectomy. Ramos et al. (18) did evaluate prediction for poor outcome with a multitude of clinical baseline parameters and CT-derived imaging features in a much larger cohort. Not including mismatch variables, their highest AUC was 0.81. Van Os et al. similarly showed a considerable improvement for prediction by adding treatment-associated variables in a study including CT imaging (19).

Interestingly, prediction for either poor or favorable outcome did not improve by including MR-mismatch variables. These results need to be interpreted together with the characteristics of the cohort: Most patients included had a target mismatch as defined in the inclusion criteria for the EXTEND-IA study (6) with an infarct core of <70 ml on ADC maps and a comparably larger volume of hypoperfused tissue on Tmax maps with a mismatch volume ratio of 1.2 or higher. While the treatment effect of MT is maintained even in patients with larger infarct cores (20), individual patients with a target mismatch still have poor outcome. Therefore, it is intriguing that poor outcome prediction in our study was possible based on clinical baseline variables only: For those patients, the potential predictive value of MR mismatch variables could be already encoded in the clinical baseline information. A similar conclusion can be drawn for patients with favorable outcome: improved prediction was much more dependent on speed and the extent of recanalization than on MR mismatch. However, this does not preclude the possibility that a target mismatch is still a valid selection criterion for patients undergoing MT. Our sample does not allow a conclusion about the potential predictive value of patients without MR mismatch. Prediction models could be improved by including patients who underwent MT without a target mismatch profile and larger infarct cores (21).

Despite these findings, the overall performance of the ML models tested in our study could be improved. Considering the potential power of ML algorithms to extract patterns, our findings suggest that important variables for outcome prediction might not be included in today's clinical decision-making. It

is conceivable that there are so far unknown or undetectable variables. This warrants further studies including more and new prediction variables and biomarkers as well as direct integration of multimodal imaging and clinical information.

A deep learning model including raw imaging data and not derived variables might extract further, previously unknown predictive information. For example, these models might be able to account for inherent errors in the definition of infarct core (22) or individual susceptibility of brain tissue (21).

Finally, our results show that the predictive value can differ significantly between two different dichotomization paradigms or different "cutoffs" (mRS 0–2 vs. 3–6 and mRS 5/6 vs. 0–4). Unnecessary dichotomization of the mRS can be suboptimal (23). Researchers and clinicians should be aware that there are relevant differences between dichotomization paradigms. Defining more accurate outcome or premorbidity scores might improve future prediction models.

Our study has some limitations. It is based on a relatively small and retrospective patient cohort. Thus, our results must be understood as an exploratory analysis for future research.

Our data reach back to 2009. This might have influenced outcome due to improvements in thrombectomy technique and accelerated workflows. However, there was no significant outcome difference between patients treated 2009–2013 vs. 2014–2018.

CONCLUSION

Our results suggest that a prediction of poor outcome (mRS of 5 or 6) after MT can be based on clinical baseline variables only. Speed and extent of thrombectomy did not seem to influence poor outcome but were important for favorable outcome (mRS 0–2). The predictive value of a target MR mismatch with smaller infarct core and larger penumbra was not relevant and could be already captured by clinical baseline variables. However, our sample does not allow a conclusion about the predictive value in patients without target MR mismatch.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Heidelberg University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MMu, AP, CW, MMö, PR, SN, and MB acquired the data and organized the database. VM, AH, EZ, and DF performed and developed the machine learning and statistical

analysis. MMu wrote the first draft of the manuscript. VM, AH, and DF wrote sections of the manuscript. All authors contributed to the conception, design of the study, manuscript revision, read, and approved the submitted version.

REFERENCES

- Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet*. (2016) 387:1723–31. doi: 10.1016/S0140-6736(16)00163-X
- Saver JL, Goyal M, van der Lugt A, Menon BK, Majoie CB, Dippel DW, et al. Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. *JAMA*. (2016) 316:1279–89. doi: 10.1001/jama.2016.13647
- Livne M, Boldsen JK, Mikkelsen IK, Fiebach JB, Sobesky J, Mouridsen K. Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke. *Stroke*. (2018) 49:912–8. doi: 10.1161/STROKEAHA.117.019440
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. (2020) 20:310. doi: 10.1186/s12911-020-01332-6
- Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. Opening the black box of artificial intelligence for clinical decision support: a study predicting stroke outcome. *PLoS ONE*. (2020) 15:e0231166. doi: 10.1371/journal.pone.0231166
- Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med*. (2015) 372:1009–18. doi: 10.1056/NEJMoa1414792
- Jayaraman MV, Grossberg JA, Meisel KM, Shaikhouni A, Silver B. The clinical and radiographic importance of distinguishing partial from near-complete reperfusion following intra-arterial stroke therapy. *Am J Neuroradiol*. (2013) 34:135–9. doi: 10.3174/ajnr.A3278
- Goyal M, Almekhlafi MA, Cognard C, McTaggart R, Blackham K, Biondi A, et al. Which patients with acute stroke due to proximal occlusion should not be treated with endovascular thrombectomy? *Neuroradiology*. (2019) 61:3–8. doi: 10.1007/s00234-018-2117-y
- Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke*. (1999) 30:1538–41. doi: 10.1161/01.STR.30.8.1538
- Zaidat OO, Yoo AJ, Khatri P, Tomsick TA, Von Kummer R, Saver JL, et al. Recommendations on angiographic revascularization grading standards for acute ischemic stroke. *Stroke*. (2013) 44:2650–63. doi: 10.1161/STROKEAHA.113.019172
- Potreck A, Loebel S, Pfaff J, Østergaard L, Mouridsen K, Radbruch A, et al. Increased volumes of mildly elevated capillary transit time heterogeneity positively predict favorable outcome and negatively predict intracranial hemorrhage in acute ischemic stroke with large vessel occlusion. *Eur Radiol*. (2019) 29:3523–32. doi: 10.1007/s00330-019-06064-4
- Purushotham A, Campbell BCV, Straka M, Mlynash M, Olivot JM, Bammer R, et al. Apparent diffusion coefficient threshold for delineation of ischemic core. *Int J Stroke*. (2015) 10:348–53. doi: 10.1111/ijss.12068
- Zaro-Weber O, Moeller-Hartmann W, Siegmund D, Kandziora A, Schuster A, Heiss WD, et al. MRI-based mismatch detection in acute ischemic stroke: optimal PWI maps and thresholds validated with PET. *J Cereb Blood Flow Metab*. (2017) 37:3176–83. doi: 10.1177/0271678X16685574
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*. Curran Associates Inc. (2017). p. 4768–77.
- Debs N, Cho TH, Rousseau D, Berthezène Y, Buisson M, Eker O, et al. Impact of the reperfusion status for predicting the final stroke infarct using deep learning. *Neuroimage Clin*. (2021) 29:102548. doi: 10.1016/j.nicl.2020.102548
- Hamann J, Herzog L, Wehrli C, Dobrocky T, Bink A, Piccirelli M, et al. Machine learning based outcome prediction in stroke patients with MCA-M1 occlusions and early thrombectomy. *Eur J Neurol*. (2021) 28:1234–243. doi: 10.1111/ene.14651
- Brugnara G, Neuberger U, Mahmutoglu MA, Foltyn M, Herweh C, Nagel S, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke*. (2020) 51:3541–51. doi: 10.1161/STROKEAHA.120.030287
- Ramos LA, Kappelhof M, van Os HJA, Chalos V, Van Kranendonk K, Kruyt ND, et al. Predicting poor outcome before endovascular treatment in patients with acute ischemic stroke. *Front Neurol*. (2020) 11:580957. doi: 10.3389/fneur.2020.580957
- van Os HJA, Ramos LA, Hilbert A, Van Leeuwen M, Van Walderveen MA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Front Neurol*. (2018) 9:784. doi: 10.3389/fneur.2018.00784
- Campbell BCV, Majoie CBLM, Albers GW, Menon BK, Yassi N, Sharma G, et al. Penumbra imaging and functional outcome in patients with anterior circulation ischaemic stroke treated with endovascular thrombectomy versus medical therapy: a meta-analysis of individual patient-level data. *Lancet Neurol*. (2019) 18:46–55. doi: 10.1016/S1474-4422(18)30314-4
- Goyal M, Menon BK, Almekhlafi MA, Demchuk A, Hill MD. The need for better data on patients with acute stroke who are not treated because of unfavorable imaging. *AJNR Am J Neuroradiol*. (2017) 38:424–5. doi: 10.3174/ajnr.A5094
- Goyal M, Ospel JM, Menon B, Almekhlafi M, Jayaraman M, Fiehler J, et al. Challenging the ischemic core concept in acute ischemic stroke imaging. *Stroke*. 51:3147–55. doi: 10.1161/STROKEAHA.120.030620
- Ganesh A, Luengo-Fernandez R, Wharton RM, Rothwell PM. Ordinal vs dichotomous analyses of modified rankin scale, 5-year outcome, and cost of stroke. *Neurology*. (2018) 91:e1951–60. doi: 10.1212/WNL.00000000000006554

Conflict of Interest: VM reported receiving personal fees from ai4medicine outside the submitted work. AH reported receiving personal fees from ai4medicine outside the submitted work. DF reported receiving grants from the European Commission Horizon2020 PRECISE4Q No. 777107, reported receiving personal fees from and holding an equity interest in ai4medicine outside the submitted work. There is no connection, commercial exploitation, transfer, or association between the projects of ai4medicine and the results presented in this work. SN received unrelated fees for consultancy from Brainomix and Boehringer Ingelheim, payment for lectures including service on speakers' bureaus from Pfizer, Medtronic, and Bayer AG. MB received unrelated grants from Siemens, grants and personal fees from Novartis, grants from Stryker, grants from DFG, personal fees from Merck, personal fees from Bayer, personal fees from Teva, grants and personal fees from Guerbet, personal fees from Boehringer, personal fees from Vascular Dynamics, personal fees from Grifols, and grants from the European Union, all outside the submitted work. MMö received unrelated Board Membership from Codman; consultancy from Medtronic, MicroVenton,

and Stryker; payment for lectures including service on speakers bureaus' from Medtronic, MicroVention, and Stryker. PR received unrelated grants for consultancy from Boehringer and lecture fees from Bayer, Boehringer Ingelheim, BMS, Daichii Sankyo, and Pfizer.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of

the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mutke, Madai, Hilbert, Zihni, Potreck, Weyland, Möhlenbruch, Heiland, Ringleb, Nagel, Bendszus and Frey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

3. Diskussion

Um die Möglichkeiten von KI-basierten Entscheidungsunterstützungssystemen im medizinischen Kontext zu nutzen, muss neben der messbaren Verbesserung des Outcomes im Vergleich zum aktuellen Goldstandard („Clinical Benefit“) die klinische Machbarkeit gegeben sein. Dementsprechend müssen Lösungen sowohl auf klinische Evidenz wie auf die Implementierung im klinischen Arbeitsablauf ausgerichtet sein.

Wie gezeigt, können bildbasierte Biomarker im Hinblick auf den wahrscheinlich eintretenden Nutzen einer Behandlung potentiell als Stratifizierungsgrundlage dienen und somit wertvolle Informationen für Behandlungsentscheidungen liefern. Hier können Informationen wie beispielsweise das funktionelle Outcome (mRS), der Behandlungserfolg, das endgültige Infarkt volumen oder das Auftreten intrakranieller Blutungen erhoben werden.

Insbesondere die Vorhersage des Behandlungserfolgs kann genutzt werden, um die Spezifität, Sensitivität und Gesamtgenauigkeit der Patientenstratifizierung für die mechanische Thrombektomie zu verbessern. Dieser Punkt ist aktuell einer der Forschungsschwerpunkte in der Schlaganfallbehandlung. Es ist dementsprechend entscheidend sowohl die Sensitivität als auch die Spezifität von Behandlungsstratifizierungsmodellen zu optimieren, um ein besseres Verständnis der Erkrankung des individuellen Patienten und eine bessere Unterstützung der Behandlungszuteilung zu ermöglichen.

Vor diesem Hintergrund entwickelten wir eine Pipeline, die auf der Grundlage routinemäßig erhobener DICOM-Bilder mittels Simulation verschiedener Randbedingungen für Ischämieereignisse gefährdete Hirnareale identifizieren kann. Bezüglich der Integration von Ergebnissen durch die von uns entwickelte Simulation zeigt die Validierungsanalyse anhand der DSC-Perfusion bei Patienten mit steno-okklusiver Erkrankung vielversprechende erste Ergebnisse für die Erkennung von gefährdeten Hirnregionen. Damit kann die vorgestellte mechanistische Simulation des Blutflusses aus der Routinebildgebung als individueller Biomarker dienen und in Zukunft potentiell die komplexe und gesundheitsschädliche Perfusionsbildgebung ersetzen. Nach weiterer Validierung kann hier der Weg zu einer individuell optimierten Einstellung des Blutdrucks, um Ischämien oder post-Infarkt-Hyperämien zu minimieren.

Der Schlaganfall ist eine komplexe Erkrankung mit einem dynamischen Verlauf. Für das Verständnis des Schlaganfalls und die damit zusammenhängenden therapeutischen Konsequenzen ist es wichtig zu verstehen, dass es eine sehr hohe interindividuelle Varianz des Untergangs von Hirngewebe gibt: die Varianz des Hirnzelltods reicht von einigen Stunden bis 17 Stunden³⁷. Dementsprechend tragen "One-size-fits-all" Behandlungsstrategien den individuellen Merkmalen (Features) der Patienten zur unzureichend Rechnung³⁸. Die Zeitfenster der Behandlung (zum Beispiel 4,5 oder 6 Stunden) wurden durch eine statistische Nutzen-Risiko-Berechnung festgelegt. Dieser Ansatz mag zu einem Nettogewinn für das Patientenkollektiv führen, aber es bedingt aufgrund der Normalverteilung ebenso, dass viele Patienten nicht behandelt werden, die von therapeutischen Maßnahmen auch außerhalb des Zeitfensters profitieren würden, und gleichzeitig Patienten behandelt werden bei denen das Interventionsrisiko den Nutzen überwiegt. Für eine individualisierte und personalisierte Präzisionsmedizin werden daher zunehmend datengetriebene mathematische Techniken und Ansätze aus der Künstlichen Intelligenz genutzt, um individualisierte Prognosen für Patienten zu erstellen³⁹.

Wir haben mit der Analyse von Daten und den vorgestellten Ergebnissen aus den eigenen Publikationen vielversprechende Ansätze zur Personalisierung aufgezeigt. Daneben ist ein wichtiges Ziel der datenbasierten Schlaganfallforschung unbekannte Prädiktoren und Biomarker in den Rohdaten der Bildgebung (DICOM) zu entdecken. Hierfür sind Deep-Learning-Lösungen geeignet, da sie gleichzeitig Merkmale extrahieren und klassifizieren können.

Es konnte gezeigt werden, dass ein Modell zur Outcome-Vorhersage (mTICI-Score) mit einem CNN-Modell auf CTA-Bildgebung eine AUC von 0,65 erreichen konnte⁴⁰. Die höchste Leistung bei der Vorhersage des funktionellen Ergebnisses, definiert als dichotomisierter mRS-Score, wurde mit einer AUC von 0,73 durch ein CNN-Modell unter Verwendung von DWI-Bildgebung erreicht⁴¹. Unter Zugrundelegung von KI-extrahierten Bildinformationen und klinischen Daten wurde mit einem CNN- und NN-Modell eine AUC von 0,75 erzielt⁴². In Anbetracht der Schlüsselrolle der Bildgebung konnten die bildbasierten Modelle bezüglich ihrer Performanz vor allem im Hinblick auf die Nutzung von Deep-Learning-Methoden bislang allerdings die in sie gesetzten hohen Erwartungen nicht erfüllen.

Genau hier setzt unsere Arbeit ein: Mit den vorgelegten Arbeiten konnten wir zeigen, dass durch die Entwicklung von ML-/DL-basierten Verfahren auf Grundlage von Patientendaten eine sinnvolle und messbar positive Patientenstratifizierung möglich ist. Die Integration von klinischen und Bilddaten in das entwickelte Modell ermöglicht, dass auf patientenindividueller Ebene die optimale Therapieentscheidung getroffen werden kann, da **vor** der Therapie vorhergesagt wird, wie sich das Outcome unterscheidet je nachdem welche Therapieoption gewählt wird.

Um die Entscheidungsfindung in einem akuten Real-World-Setting effektiv zu unterstützen, müssen die entwickelten Modelle, Algorithmen und Lösungen Anforderungen im Hinblick auf Nutzer- und Gebrauchstauglichkeit erfüllen. So werden Lösungen und Werkzeuge, die Variablen und Daten verwenden, die nicht routinemäßig erhoben werden, die Integration in die klinischen Abläufe und die Akzeptanz durch das medizinische Personal stark beeinträchtigen. Viele entwickelte Systeme, die publiziert sind, verwenden mehr als 10 Variablen^{41,43} und führen dazu, dass Entscheidungssysteme, die auf diesen Input-Features angewiesen sind, nur ein Produkt für spezialisierte Zentren sein können – für die Anwendung in der Breite – für die ein Entscheidungsunterstützungssystem den eigentlichen Vorteil hat – sind diese Lösungen nicht praktikabel, da deren Algorithmen mehr als routinemäßig erhobene Daten zu einer prädiktiven Modellierung benötigen. Dieser sogenannte „Datenhunger“ könnte dadurch entstehen, dass die Forschung nicht mit anderen Stakeholdern ins Gespräch kommt, um standardisierte Anforderungen zu entwickeln, sondern forschungsspezifisch hochinteressante Fragestellungen bearbeiten will und die Anwendung in der Praxis als nachrangig betrachtet.

Mit den von uns entwickelten Frameworks und Systemen sind wir in der Lage klinisch-pragmatische Lösungen anzubieten, die den Input klinischer Features reduzieren und auch in der Analyse der Bildgebung minimale Anforderungen an Sequenzen und Wichtungen stellen. Unsere Modelle sind auf den routinemäßig erhobenen Daten trainiert, getestet und validiert worden sodass sie im klinischen Alltag unmittelbar ohne zusätzliche Anforderungen eingesetzt werden könnten. Darüber hinaus sollten interdisziplinäre Anstrengungen unternommen werden, um sowohl durch Nachweis der klinischen Evidenz und eine pragmatische Herangehensweise bezüglich der Integration in den klinischen Ablauf den Nutzen zukünftiger KI-basierter Entscheidungsunterstützungssysteme zu verbessern. Genau dies wird aktuell als Fortsetzung der beschriebenen Forschungsarbeiten unternommen. Ein auf den entwickelten Modellen basierendes Entscheidungsunterstützungssystem wird aktuell in 3 klinischen Zentren getestet.

Weiterhin ist es im Fortschritt der KI-Technologien insgesamt – und vor allem im Bereich der Medizin – von entscheidender Bedeutung, dass Forschung transparent und verantwortungsvoll durchgeführt wird und die Ergebnisse einen sorgfältigen Validierungsprozess durchlaufen. In der Validierung der klinisch relevanten Ergebnisse ist die Replizierbarkeit wichtig, insbesondere wenn es um prospektive Generalisierung und die Implementierung der Modelle geht. Daher ist die Identifikation und Auswahl der Parameter und Hyperparameter sorgfältig durchzuführen, da diese die Leistung und Ergebnisse der Machine und Deep Learning stark beeinflussen und eine Voreingenommenheit der Modelle und

Algorithmen, einen sog. Bias verursachen können.

Bei der Implementierung der Modelle in technologische Lösungen und Anwendungen muss neben der klinischen und theoretischen Validierung eine umfassende Überprüfung der Robustheit der Daten durchgeführt werden, insbesondere auf Overfitting und Splitting der Datensätze, damit die Validierung die realen Szenarien optimal abbildet, zum Beispiel durch Cross Validation⁴⁴. Schließlich ist es für die Zukunft von essentieller Bedeutung die Abhängigkeit der künstlichen Intelligenz von gelabelten, d.h. von Experten markierten Daten zu reduzieren, um generalisierungsfähige und robuste Modelle zu erhalten^{45,46}. Die zunehmende Sammlung von Daten in hoher Qualität und Quantität kann neben klassischen randomisierten klinischen Studien sowohl eine Verbesserung in der Modellentwicklung wie auch in der Validierung der Algorithmen dienen.

Darüber hinaus konnte mit dieser Arbeit gezeigt werden, dass für die Verbreiterung der Datenbasis die Synthese und Generierung künstlicher Daten ein effizienter Weg sein kann. Diese KI-generierten neuen Patientendaten sind nicht erfunden, sondern stellen vielmehr eine ausreichend verzerrte, verrauschte Wiedergabe der ursprünglichen Bilder dar. Wir konnten bezüglich der generierten MR-Bilder zeigen, dass eine Balance zwischen Privatheit der Daten – d.h. dass die generierten Daten keinen Rückschluss auf die Originaldaten zulassen – und der Nützlichkeit der Daten – d.h. dass die Daten noch ausreichend repräsentativ sind, um KI-performante Modelle zu trainieren – gefunden werden kann. Es muss in diesem Kontext streng berücksichtigt werden, dass eine Re-Identifizierung der Originaldaten auf jeweilige Patientenidentitäten vermieden werden muss, da dies ansonsten geltende Datenschutzbestimmungen unterlaufen würde.

Black Box

Wie beschrieben, ist für die Entwicklung und Verwendung von KI-Modellen – im Gegensatz zu anderen Segmenten – im Medizinbereich ein besonders hohes Maß an Sicherheit und Validität notwendig, da das Leben und die Gesundheit von Patienten geschützt, erhalten und wiederhergestellt werden müssen⁴⁷. Ein Kriterium ist die Erklärbarkeit (Explainability), bzw. Interpretierbarkeit und Transparenz⁴⁸. Interpretierbarkeit wird als methodische Erklärbarkeit definiert (Gewichte eines linearen Regressionsalgorithmus) im Gegensatz zu Verständlichkeit, die eine symbolische Darstellung einer Ausgabe ist⁴⁹. Die Darstellung von Erklärbarkeit und deren Visualisierung für die Nutzer sind von entscheidender Bedeutung, da ein Hauptziel der Entwicklung von KI-Modellen deren Integration in klinische Entscheidungsunterstützungssysteme (CDSS) ist^{14,15}.

Diese Anforderungen an Erklärbarkeit und Interpretierbarkeit wurde in der vorgelegten Arbeit berücksichtigt: Wir konnten die Bedeutung der klinischen Merkmale mit mehreren Methoden (Deep Taylor Decomposition, Shapley-Werten, Modellkoeffizienten) darstellen und in eine Reihenfolge („Importance Ranking“) bringen. Dies kann in Bezug auf Akzeptanz und Nachvollziehbarkeit durch das behandelnde medizinische Personal ein entscheidendes Kriterium sein^{49,50}. Die Leitlinien für klinische Unterstützungssysteme sollten daher eine gute Balance zwischen Erklärbarkeit der Algorithmen auf der einen Seite und der inhärenten Unmöglichkeit von Deep Learning Ansätzen jeden Schritt nachvollziehbar zu machen, finden. Vor dem Hintergrund, dass Ärztinnen und Ärzte und weiteres medizinisches Personal auch in anderen Feldern ähnliche Unsicherheiten in der medizinischen Entscheidungsfindung akzeptiert haben und nicht jeden Prozess zu 100 Prozent nachvollziehen und verstehen müssen, sollte die Erklärbarkeit ein unterstützendes Instrument sein.

Entscheidend ist, dass eine klare und umfassende klinische Validierung erfolgt und der Benefit für Patienten bewiesen wird. Darüber hinaus ist dieser Punkt, die Öffnung der Black Box ist auch ein entscheidender Aspekt in den regulatorischen Anforderungen und Regelungen in MDR und FDA⁵¹.

In Gesamtschau der in den einzelnen Arbeiten genannten Limitationen können zusammenfassend die folgenden Punkte genannt werden: Trotz hoher Qualität gab es hinsichtlich der Menge der Trainingsdaten eine Einschränkung. Insgesamt ist, wie beschrieben, in der medizinischen Forschung die Verfügbarkeit von Daten aufgrund Datensicherheit, Privacy und dem Erfordernis des Informed Consent im Vergleich zu anderen ML/DL Use Cases erheblich reduziert. Daneben ist zu betonen, dass die generierten Modelle und Algorithmen nicht ohne weiteres auf „neue“ Patientenkollektive angewandt werden können. Bedingt durch die Datenquellen, die zwar multizentrisch waren, jedoch in ihrer Heterogenität eingeschränkt, muss vor Anwendung auf andere Patienten eine Anpassung und Feinjustierung der Modelle erfolgen, um den inhärenten Bias zu reduzieren. Im Bereich der Analyse der Bilddaten gab es Restriktionen in Bezug auf Rechenkapazität und Rechenleistung der verwendeten Systeme. Dies führte zur Priorisierung von Ansätzen und Methodik. Schließlich wurden die vorgestellten Ergebnisse auf Grundlage von Patientendaten mit cerebrovaskulären Erkrankungen erarbeitet. Eine Erweiterung der Indikationen ist sinnvoll, bedarf großer Ressourcen und wird auf Grundlage der vorgestellten Ergebnisse momentan initiiert und teilweise schon durchgeführt.

4. Zusammenfassung

Mit dieser Arbeit konnte gezeigt werden, dass eine Anwendung von KI-basierten Ansätzen in der Schlaganfallbehandlung zu einer Personalisierung der Therapiestrategie und einer Verbesserung des Outcomes führen kann. Wir haben erstmals in großem Umfang multidimensionale Daten für die Entwicklung KI-basierter Modelle zur Verbesserung der Akutbehandlung des Schlaganfalls nutzbar gemacht und auf deren Grundlage reliable, akkurate und nachvollziehbare Algorithmen entwickelt.

Zum einen konnten wir ein Framework zur Simulation der Hämodynamik der Hirngefäße entwickeln und hiermit erstmals für zerebrovaskulären Erkrankungen einen simulationsbasierten Ansatz im Sinne einer Präzisionsmedizin konzipieren und entwickeln. Darüber hinaus konnten wir erfolgreich Daten mittels generativen Ansätzen, hier mit Generative Adversarial Networks (GANs), synthetisieren, um der Herausforderung der Datenknappheit für die Entwicklung von KI-Modellen insbesondere im Bereich von Bilddaten zu begegnen. Mit der abschließenden Arbeit konnte gezeigt werden, dass die Zusammenführung dieser wissenschaftlichen Erkenntnisse als Grundlage für eine Machine Learning-basierte Entscheidungsunterstützung zur Behandlung des akuten Schlaganfalls dienen kann.

Für die Anwendung einer KI-basierten Entscheidungsunterstützung sind in der Praxis noch verschiedene Validierungsschritte in Bezug auf den medizinischen Nutzen und die Implementierung in den klinischen Workflow zu durchlaufen. Insbesondere muss klar gezeigt werden, dass durch die Anwendung der Methoden und Modelle ein messbarer klinischer Nutzen für die Patienten entsteht (Outcome-Verbesserung).

Insgesamt konnte durch die Entwicklung von KI-Modellen und Algorithmen somit der Grundstein für eine Therapieunterstützung auf Basis der individuellen Features, beziehungsweise Charakteristika des Individuums erfolgreich gelegt werden. Damit ist für ein klinisches Unterstützungssystem für die Behandlung des akuten Schlaganfalls, welches Ärztinnen und Ärzte durch zusätzliche Informationen eine personalisierte und damit bessere Therapiestratifizierung ermöglicht, der Grundstein gelegt worden. Dass die vorgelegte Arbeit eine große Reichweite und einen wissenschaftlichen Impact hat, zeigt die Tatsache, dass die Weiterentwicklung, Validierung und Implementierung der entwickelten Modelle und Algorithmen momentan schon in weiteren Projekten praktisch durchgeführt wird.

5. Literaturangaben

1. GBD 2015 Neurological Disorders Collaborator Group. Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Neurol.* 2017 Nov;16(11):877–97.)
2. Li KHC, Jesuthasan A, Kui C, Davies R, Tse G, Lip GYH. Acute ischemic stroke management: concepts and controversies. A narrative review. *Expert Rev Neurother.* 2021 Jan 2;21(1):65–79.
3. Berge E, Whiteley W, Audebert H, De Marchis G, Fonseca AC, Padiglioni C, et al. European Stroke Organisation (ESO) guidelines on intravenous thrombolysis for acute ischaemic stroke. *Eur Stroke J.* 2021 Mar 1;6(1):I–LXII.
4. Turc G, Bhogal P, Fischer U, Khatri P, Lobotesis K, Mazighi M, 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 Mar 1;4(1):6–12.)
5. Jamieson T, Goldfarb A. Clinical considerations when applying machine learning to decision-support tasks versus automation. *BMJ Qual Saf.* 2019 Oct 1;28(10):778–81.
6. Nagel S, Sinha D, Day D, Reith W, Chapot R, Papanagiotou P, et al. e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. *Int J Stroke.* 2017 Aug;12(6):615–22.
7. Brinjikji W, Abbasi M, Arnold C, Benson JC, Braksick SA, Campeau N, et al. e-ASPECTS software improves interobserver agreement and accuracy of interpretation of aspects score. *Interv Neuroradiol.* 2021 Dec 1;27(6):781–7.
8. Pfaff J, Herweh C, Schieber S, Schönenberger S, Bösel J, Ringleb PA, et al. e-ASPECTS Correlates with and Is Predictive of Outcome after Mechanical Thrombectomy. *AJNR Am J Neuroradiol.* 2017 Aug;38(8):1594–
9. Maegerlein C, Fischer J, Mönch S, Berndt M, Wunderlich S, Seifert CL, et al. Automated Calculation of the Alberta Stroke Program Early CT Score: Feasibility and Reliability. *Radiology.* 2019 Apr;291(1):141–8.)
10. Grunwald IQ, Kulikovski J, Reith W, Gerry S, Namias R, Politi M, et al. Collateral Automation for Triage in Stroke: Evaluating Automated Scoring of Collaterals in Acute Stroke on Computed Tomography Scans. *Cerebrovasc Dis.* 2019;47(5–6):217–22.

11. Kellner E, Reiser M, Kiselev VG, Maurer CJ, Beume LA, Urbach H, et al. Automated Infarct Core Volumetry Within the Hypoperfused Tissue: Technical Implementation and Evaluation. *J Comput Assist Tomogr.* 2017 Aug;41(4):515–20.
12. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. *N Engl J Med.* 2023 May 25;388(21):1981-1990. doi: 10.1056/NEJMra23017255.
13. Khamparia A, Singh KM. A systematic review on deep learning architectures and applications. *Expert Syst.* 2019; 36: e12400. <https://doi.org/10.1111/exsy.12400>.
14. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* 2019; 28: 231–237. <https://doi.org/10.1136/bmjqs-2018-008370> PMID: 30636200;
15. Ashrafian H, Darzi A. Transforming health policy through machine learning. *PLOS Med.* 2018; 15: e1002692. <https://doi.org/10.1371/journal.pmed.1002692> PMID: 30422977
16. Griffiths F, Ooi M, *IEEE Instrum. Meas. Mag.* 2018, 21, 29
17. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med.* 2019; 25: 24–29. <https://doi.org/10.1038/s41591-018-0316-z> PMID: 30617335;
18. Leshno M, Lin V, Pinkus A, Schocken S, *Neural Netw.* 1993, 6, 861
19. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O, written on behalf of AMEB-DCTCG (2019) Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med* 7:152.
20. Livne M, Boldsen JK, Mikkelsen IK, Fiebach JB, Sobesky J, Mouridsen K (2018) Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke. *Stroke* 49:912–918)
21. Higgins D, Madai VI. (2020), From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Adv. Intell. Syst.*, 2: 2000052. <https://doi.org/10.1002/aisy.202000052>
22. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.* 2018; 19: 1236–1246. <https://doi.org/10.1093/bib/bbx044> PMID: 28481991

23. Luo L, Li J, Liu C, Shen W. Using machine-learning methods to support health-care professionals in making admission decisions. *Int J Health Plann Manage*. 2019; 34: e1236–e1246. <https://doi.org/10.1002/hpm.2769> PMID: 30957270;
24. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang S-W, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLOS ONE*. 2019; 14: e0221202. <https://doi.org/10.1371/journal.pone.0221202> PMID: 31442238;
25. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018; 6: 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
26. Nathans LL, Oswald FL, Nimon K. Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. 2012; 17: 19.
27. Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. *Digit Signal Process*. 2018; 73: 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
28. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, NY, USA: ACM; 2018. pp. 559–560. <https://doi.org/10.1145/3233547.3233667>
29. Shortliffe EH, Sepulveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*. 2018; 320: 2199–2200. <https://doi.org/10.1001/jama.2018.17163> PMID: 30398550
30. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLOS Med*. 2018; 15: e1002689. <https://doi.org/10.1371/journal.pmed.1002689> PMID: 30399149
31. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017. pp. 4765–4774. Available: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
32. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, **Frey D**. Opening the Black Box of Artificial Intelligence for Clinical Decision Support: A Study Predicting Stroke Outcome. *PlosOne*, April 2020(4), e02311662 <https://doi.org/10.1371/journal.pone.0231166>
33. Frey D, Livne M, Leppin H, Akay EM, Aydin OU, Behland J, Sobesky J, Vajkoczy P, Madai VI. A precision medicine framework for personalized simulation of hemodynamics in cerebrovascular disease. *Biomed Eng Online*. 2021 May 1;20(1):44. doi: 10.1186/s12938-021-00880-w. PMID: 33933080; PMCID: PMC8088619

34. Kossen T, Subramaniam P, Madai VI, Hennemuth A, Hildebrand K, Hilbert A, Sobesky J, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput. Biol. Med.* (2021). <https://doi.org/10.1016/j.combiomed.2021.104254>
35. Kossen T, Hirzel MA, Madai VI, Boenisch F, Hennemuth A, Hildebrand K, Pokutta S, Sharma K, Hilbert A, Sobesky J, Galinovic I, Khalil AA, Fiebach JB, Frey D. Toward Sharing Brain Images: Differentially Private TOF-MRA Images With Segmentation Labels Using Generative Adversarial Networks. *Frontiers in Artificial Intelligence* 5 (May 2, 2022): 813842. <https://doi.org/10.3389/frai.2022.813842>
36. Mutke MA, Madai VA, Hilbert A, Zihni E, Potreck A, Weyland CS, Möhlenbruch MA, Heiland S, Ringleb PA, Nagel S, Bendszus M, Frey D. Comparing poor and favorable outcome prediction with machine learning after mechanical thrombectomy in acute ischemic stroke. *Frontiers in Neuroscience*. 2022 <https://doi.org/10.3389/fneur.2022.737667>
37. Marchal G, Beaudouin V, Rioux P, de la Sayette V, Doze FL, Viader F, et al. Prolonged persistence of substantial volumes of potentially viable brain tissue after stroke a correlative PET-CT Study With Voxel-Based Data Analysis. *Stroke*. 1996;27:599–606. <https://doi.org/10.1161/01.STR.27.4.59>
38. Powers WJ, Rabinstein AA, Teri A, Adeoye OM, Bambakidis NC, Kyra B, 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:e46–99. <https://doi.org/10.1161/STR.00000000000000158>.
39. Rostanski SK, Marshall RS. Precision medicine for ischemic stroke. *JAMA Neurol*. 2016;73:773–4. <https://doi.org/10.1001/jama.neuro.2016.0087>
40. Hilbert A, Ramos LA, van Os HJA, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput. Biol. Med.*, vol. 115, p. 103516, Dec. 2019, doi: 10.1016/j.combiomed.2019.103516
41. Nishi H, Oishi N, Ishii A. et al. Deep Learning–Derived High-Level Neuroimaging Features Predict Clinical Outcomes for Large Vessel Occlusion,” *Stroke*, vol. 51, no. 5, pp. 1484–1492, May 2020, doi: 10.1161/STROKEAHA.119.028101.
42. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes. *Acad. Radiol.*, vol. 27, no. 2, pp. e19–e23, Feb. 2020, doi: 10.1016/j.acra.2019.03.015.

43. Ho KC, Speier W, El-Saden S, Arnold CW. Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features. *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2017, pp. 892–901, 2017
44. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, vol. 145, pp. 166–179, Jan. 2017, doi: 10.1016/j.neuroimage.2016.10.038.
45. Feng R, Badgeley M, Mocco J, Oermann EK. Deep learning guided stroke management: a review of clinical applications. *J. NeuroInterventional Surg.*, vol. 10, no. 4, pp. 358–362, Apr. 2018, doi: 10.1136/neurintsurg-2017-013355.
46. Eitel F, Schulz MA, Seiler M, Walter H, Ritter K. Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. *Exp. Neurol.*, vol. 339, p. 113608, May 2021, doi: 10.1016/j.expneurol.2021.113608.
47. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf.* 2019; 28: 238–241. <https://doi.org/10.1136/bmjqs-2018-008551> PMID: 30291179
48. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable Machine Learning for Scientific Insights and Discoveries. *ArXiv190508883 Cs Stat.* 2019 [cited 3 Sep 2019]. Available: <http://arxiv.org/abs/1905.08883>
49. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *ArXiv171000794 Cs.* 2017 [cited 2 Aug 2022]. Available: <http://arxiv.org/abs/1710.00794>
50. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep.* 2019; 49: 15–21. <https://doi.org/10.1002/hast.973> PMID: 30790315
51. johner-institut/ai-guideline. In: GitHub [Internet]. [cited 2 Aug 2022]. Available: <https://github.com/johnerinstitut/ai-guideline>

6. Abkürzungsverzeichnis

ACA	=	Anterior cerebral artery	NB	=	Naive Bayes
AI	=	Artificial Intelligence	NIHSS	=	National Institutes of Health Stroke Scale
AIS	=	Acute ischemic stroke	PCA	=	Posterior cerebral artery
AMNA	=	Adjusted modified nodal analysis	ROC	=	Receiver-operating-characteristic
ANN	=	Artificial neural nets	SVM	=	Support Vector Machine
ASPECTS	=	Alberta Stroke Program Early CT Score	SVMC	=	Support Vector Machine Classifier
AUC	=	Area-under-the-curve	SHAP	=	Shapley Additive Explanations
CBF	=	Cerebral blood flow			Transient/transitory ischemic attack
CDSS	=	Clinical Decision Support Systems	TIA	=	
CT	=	Computer Tomography	Tmax	=	Time-to-maximum
CTA	=	Computer Tomography Angiography	TOF	=	Time of flight
CoW	=	Circle of Willis	TTP	=	Time to peak
DICOM	=	Digital Imaging and Communications in Medicine	VIF	=	Variance inflation factor
DSC	=	Dice Similarity Coefficient	WGAN	=	Wasserstein GAN
DWI	=	Diffusion weighted imaging	WGAN-GP	=	Wasserstein GAN with gradient penalty
EC-IC	=	Extracranial-Intracranial	WGAN-GP-SN	=	Wasserstein GAN with gradient penalty and spectral normalisation
FDA	=	Food and Drug Administration	xAI	=	Explainable Artificial Intelligence
FID	=	Fréchet Inception Distance			
GAN	=	Generative Adversarial Networks			
GLM	=	Generalized linear model			
GUI	=	Graphical user interface			
HD	=	Hausdorff-Distanz			
KI	=	Künstliche Intelligenz			
LVO	=	Large vessel occlusion			
MCA	=	Middle cerebral artery			
MDR	=	Medical Device Regulation			
ML	=	Machine Learning			
MLP	=	Multilayer Perceptron			
mRS	=	Modified Rankin Scale			
MNA	=	Modified nodal analysis			
MT	=	Mechanical thrombectomy			
MRI	=	Magnetic resonance imaging			
MRA	=	Magnetic resonance			
MTT	=	angiography			
		Mean transit time			
mTICI	=	Modified treatment in cerebral ischemia			

7. Danksagung

Hiermit danke ich Prof. Dr. med. Peter Vajkoczy für seine große Unterstützung, Motivation und Förderung. Die Grundlagen dieser Arbeit wurden durch vielfache Diskussionen und seinen konstruktiven und kontinuierlichen Support gelegt. Er hat mir den Freiraum gegeben, eine neue Dimension zu sehen, um innovative Technologie für die medizinische Forschung und Versorgung zu entwickeln und einzusetzen.

Daneben möchte ich mich bei Dr. Vince Madai bedanken ohne dessen kompetente, dauerhafte Unterstützung die Erstellung einer hoch innovativen Arbeit in einem neu geschaffenen Feld nicht möglich gewesen wäre. Insgesamt gilt mein Dank allen aktuellen und ehemaligen Mitgliedern des Charité Lab for AI in Medicine, deren harte und herausfordernde Arbeit es möglich gemacht haben, Grenzen zu durchbrechen und neue Methoden zum Wohle der zukünftigen Patienten verantwortungsvoll und effektiv einzusetzen.

Mein größter Dank gilt meiner Familie: Yvonne, ohne deren großartige und unermüdliche Unterstützung und konstruktive Begleitung diese Arbeit nicht möglich gewesen wäre. Sie ist meine Inspiration!

Und meinen 3 wundervollen Kindern Pepe, Mimi und Olivia, die trotz vieler Arbeit und Herausforderungen mich immer unterstützt und ermutigt haben. Tausend Dank Euch allen!

8. Erklärung

Hiermit erkläre ich, dass - weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde, - die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden, - mir die geltende Habilitationsordnung bekannt ist. Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

Datum 23.06.2023

Unterschrift