Neurosurgical Review (2021) 44:2837–2846
https://doi.org/10.1007/s10143-020-01453-6

ORIGINAL ARTICLE

Outcome prediction in aneurysmal subarachnoid hemorrhage: a comparison of machine learning methods and established clinico-radiological scores

Nora Franziska Dengler[1] · Vince Istvan Madai[2,3] · Meike Unteroberdörster[1] · Esra Zihni[2,4] · Sophie Charlotte Brune[1] · Adam Hilbert[2] · Michelle Livne[2] · Stefan Wolf[1] · Peter Vajkoczy[1] · Dietmar Frey[1,2]

Received: 9 July 2020 / Revised: 11 November 2020 / Accepted: 1 December 2020 / Published online: 20 January 2021

Abstract

Reliable prediction of outcomes of aneurysmal subarachnoid hemorrhage (aSAH) based on factors available at patient admission may support responsible allocation of resources as well as treatment decisions. Radiographic and clinical scoring systems may help clinicians estimate disease severity, but their predictive value is limited, especially in devising treatment strategies. In this study, we aimed to examine whether a machine learning (ML) approach using variables available on admission may improve outcome prediction in aSAH compared to established scoring systems. Combined clinical and radiographic features as well as standard scores (Hunt & Hess, WFNS, BNI, Fisher, and VASOGRADE) available on patient admission were analyzed using a consecutive single-center database of patients that presented with aSAH (n = 388). Different ML models (seven algorithms including three types of traditional generalized linear models, as well as a tree bosting algorithm, a support vector machine classifier (SVMC), a Naive Bayes (NB) classifier, and a multilayer perceptron (MLP) artificial neural net) were trained for single features, scores, and combined features with a random split into training and test sets (4:1 ratio), ten-fold cross-validation, and 50 shuffles. For combined features, feature importance was calculated. There was no difference in performance between traditional and other ML applications using traditional clinico-radiographic features. Also, no relevant difference was identified between a combined set of clinico-radiological features available on admission (highest AUC 0.78, tree boosting) and the best performing clinical score GCS (highest AUC 0.76, tree boosting). GCS and age were the most important variables for the feature combination. In this cohort of patients with aSAH, the performance of functional outcome prediction by machine learning techniques was comparable to traditional methods and established clinical scores. Future work is necessary to examine input variables other than traditional clinico-radiographic features and to evaluate whether a higher performance for outcome prediction in aSAH can be achieved.

Keywords Aneurysmal subarachnoid hemorrhage · Outcome prediction · Deep learning · Artificial neural net · Tree boosting

✉ Nora Franziska Dengler
Nora.Dengler@charite.de

[1] Department of Neurosurgery, Charité Universitaetsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany

[2] CLAIM – Charité Lab for AI in Medicine, Charité Universitaetsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

[3] School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, 15 Bartholomew Row, Birmingham B5 5JU, UK

[4] Technological University Dublin, Aungier St, Dublin D02 HW71, Ireland

Abbreviations

ANN	Artificial neural net
aSAH	Aneurysmal subarachnoid hemorrhage
AUC	Area under the curve
aVS	Angiographic vasospasm,
BNI	Barrow Neurological Institute scale
CI	Cerebral infarction
CT	Computed tomography
DIND	Delayed ischemic neurological deficit
DL	Deep learning
DSA	Digital subtraction angiogram
GCS	Glasgow Coma Scale
GLM	Generalized linear model
ICH	Intracerebral hemorrhage

ICU	Intensive care unit
IVH	Intraventricular hemorrhage
ML	Machine learning
MLP	Multilayer perceptron artificial neural net
mRS	Modified Rankin Scale
NB	Naive Bayes
ROC	Receiver operating characteristics
SVMC	Support vector machine classifier
WFNS	World Federation of Neurological Societies

Introduction

Scoring systems help clinicians to classify the severity of a disease, to estimate the natural course, and to select treatment strategies[1, 23]. For aneurysmal subarachnoid hemorrhage (aSAH), the Hunt and Hess scale and the WFNS scale have been used in clinical routine for many decades. Both scores are based on clinical patient characteristics in terms of consciousness and neurological deficits [15, 28]. Numerous radiographic scores were introduced, using qualitative imaging features like the dispersion of the subarachnoid blood clot as well as the presence of intraventricular hemorrhage (IVH) or intracerebral hemorrhage (ICH)[13, 14]. The first semi-quantitative radiological predictive tool was proposed by the Barrow Neurological Institute (BNI) in 2012 [35]. However, to date, neither clinical nor radiographic scores reached the accuracy needed for definite decision-making [9, 34]. Combinations of radiographic and clinical features using traditional statistic methods have also not resulted in improved predictions [6, 17].

There is a clinical need to find tools that facilitate individualized risk stratification at an early time point of the disease to responsibly allocate resources (e.g., intensive care unit (ICU) beds) and decide on treatment strategies. Recently, machine learning (ML) approaches are increasingly applied in healthcare. Such techniques include support vector machines, decision trees, Bayesian approaches, and artificial neural networks. They may improve the clinical performance of predictive models [27, 30]. In this context, especially artificial neural nets (ANN) and methods of tree boosting, a decision tree–based algorithm, showed better performance than traditional ML approaches such as linear and logistic regression for numerous applications [12, 18, 37]. However, the substantial heterogeneity of clinical questions, input and output variables, and applied algorithms may reduce traceability and reproducibility [24, 32].

We aimed to examine in this study whether applying ML techniques improves the performance of outcome prediction in aSAH. First, we analyzed whether existing scores would benefit from the application of ML techniques. Second, we combined a set of traditional clinico-radiological features that showed to be relevant for patient outcome with availability on admission and compared its predictive performance to traditional clinical scores to maintain transparency and comparability with existing studies.

Materials and methods

Data collection

We included radiographic and clinical data of consecutive patients after aSAH treated at two hospitals of a single academic institution between 2009 and 2015. The study was approved by the ethics review board of Charité Universitaetsmedizin Berlin (EA1/291/14). Patients with documented aSAH on CT or positive lumbar puncture were enrolled in the study. Patients with bleeding sources other than an intracranial aneurysm documented by CT angiogram or digital subtraction angiography were excluded. Clinical scores were applied on admission and radiographic scores were calculated based on admission CT.

Patient management

The local treatment protocol was previously described [8, 25]. In brief, patients were treated according to international guidelines with early aneurysm occlusion, clinical and/or multimodal invasive neuromonitoring in the ICU [2].

Outcome assessment

The primary outcome measure in our study was functional outcome using the Modified Rankin Scale (mRS) [31]. Clinical outcome was acquired from files during scheduled control visits 6–12 months after the initial hemorrhage. If sufficient information was not available for mRS determination, a systematic telephone interview was conducted. Both assessments were blinded to initial SAH severity grading. Outcome was dichotomized as favorable (mRS 0–2) or unfavorable (mRS 3–6).

Scores

CT, clinical, and combined scores were applied according to the respective literature [13–15, 28, 35]. A routine assessment of Hunt and Hess grading, neurological deficits, and GCS was performed prospectively on admission and electronically documented. Calculation of WFNS score was therefore indirectly possible based on GCS. Radiographic data were retrospectively assessed by an experienced neurosurgeon blinded for outcome. VASOGRADE was calculated based on this retrospective and prospective data assessment and according to previous literature [3]. Moreover, clinical data assessment included patient age, sex, and pupillary state (equal, reactive to light vs.

fixation of one or more pupil). Additional radiographic features that were included were presence of ICH, IVH, subdural hematoma (SDH), and midline shift (MLS) larger than 5 mm. Aneurysm size and aneurysm position dichotomized for posterior or anterior location were assessed with the help of CT angiography and/or digital subtraction angiogram (DSA) on admission. An overview of the scores used for prediction is presented in Table 1.

Feature selection

The available database consisted of 408 patients. Of these, 20 patients did not have mRS values and were thus excluded resulting in the final number of 388 patients. There were very few missing values present (age 0.8%, ICH 0.3%, MLS 0.5%, SDH 0.3%, localization 0.8%, VASOGRADE score 1%). We used mean/mode imputation in each fold to impute missing values (see section “Model training and validation”).

For input features, inclusion criteria were a ratio of at least 1 to 4 for binary variables (absence/presence) and no more than 10% missing values. As an exception, we included pupil status (13.4 % of patients with pathological pupil status) due to its clinical importance (20). The following features were available: age, sex, pupil status, presence of IVH, presence of ICH, presence of MLS, presence of SDH, and the localization of the aneurysm. Categorical features with more than two or more categories were transformed into binary features as they had too few instances per category. Pupil status was dichotomized to “both pupils reactive to light” vs “pathological.” Radiologically defined ICH was dichotomized to “yes”/”no.” Radiologically defined change in the brain midline was dichotomized to shift > 5 mm “yes”/”no.” Location of the aneurysm was dichotomized to anterior circulation “yes”/”no.” Thus, all resulting features were either binary categories or continuous.

Table 1 Overview of the 8 different feature selections. ICH = intracranial hemorrhage; IVH = intraventricular hemorrhage; SDH = subdural hemorrhage; GCS = Glasgow Coma Scale score; BNI = Barrow Neurological Institute scale; WFNS = World Federation of Neurosurgical Societies

	Feature(s)
1	Hunt and Hess Score value
2	WFNS score value
3	Original Fisher score value
4	Modified Fisher score value
5	VASOGRADE score value
6	BNI score value
7	Glasgow Coma Scale value
8	Age, GCS score, sex, pupil status, presence of IVH, presence of ICH, presence of midline shift > 5 mm, presence of SDH, localization anterior circulation or other, BNI score

Model selection

We trained a model for each single score (HH, WFNS, original Fisher, modified Fisher, VASOGRADE combined, BNI, and GCS). Additionally, we a priori constructed a combined feature set of selected scores (GCS, BNI) and individual features in a way that all clinically relevant (age, pupil state, and GCS) and radiographically important parameters (including IVH, ICH, SDH, MLS, and BNI for semi-quantitative description of the thickness of subarachnoidal blood) available on admission were included. The final set of input features for each tested model is listed in Table 1.

Machine learning framework

The ML framework was written in Python using standard ML libraries. The main framework has previously been described in full technical detail in an open access publication [38]. The current framework code is available on GitHub (https://github.com/prediction2020/explainable-predictive-models). In a supervised ML approach, the above-mentioned clinical parameters and clinical scores (see also Table 1) were used to predict the final outcome of aSAH patients according to mRS. The applied dichotomization resulted in 181 positive (favorable outcome) and 207 negative (unfavorable outcome) cases. This small imbalance causes negligible bias and therefore did not warrant a sub-sampling approach limiting the available data for model training.

Applied algorithms

Seven different algorithms were applied for all eight feature selections. We used three types of generalized linear models (GLM): a plain GLM, an L1 regularized GLM (equivalent to Lasso logistic regression), and a GLM elastic net adding an additional L2 regularization. Additionally, the CatBoost tree boosting algorithm, a support vector machine classifier (SVMC), a Naive Bayes (NB) classifier, and a multilayer perceptron (MLP) artificial neural net were used. For feature selection 8 (the only model with more than one feature, see also Table 1), feature importance ratings were calculated, for all seven algorithms, using SHapley Additive exPlanations (SHAP) values. A full technical overview of the algorithms and the feature importance calculations are available in the open access publication of the applied framework [38] and the GitHub page of our framework (see above). Since multicollinearity may confound the predictive performance, we estimated multicollinearity of the features using the variance inflation factor (VIF) [22].

Model 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. The models were then tuned using 10-fold cross-validation. The whole process was repeated in 50 shuffles.

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) as the primary measure. Additional performance measures were accuracy, average class accuracy, precision, recall, f1 score, negative predictive value, and specificity. To estimate calibration of the models, the Brier score was calculated. All measures are given as the median over 50 shuffles.

Interpretability assessment

The absolute values of the calculated feature importance scores were scaled to unit norm to provide comparable feature rating across models: for each of the 50 shuffles, the calculated importance scores were rescaled to the range [0,1] with their sum equal to one. Then, for each feature, the mean and standard deviation over the shuffles were calculated and reported as the final rating measures.

Results

Patient characteristics and importance of features

Three hundred eighty-eight patients with a median age of 54 years (IQR 46; 63) and a female:male ratio of 2.3 were included in the final analysis. Clinical and radiographic patient characteristics are depicted in Table 2. Functional outcome was evaluated after a median of 10 months (IQR 6; 17).

The chosen features in feature selection eight demonstrated negligible multicollinearity with VIFs < 3.48 (age 1.08, GCS 1.81, sex 2.77, pupil status 1.44, presence of IVH 2.41, presence of ICH 2.17, presence of MLS 1.85, presence of SDH 1.17, localization of the aneurysm 3.47, BNI score 1.24).

Predictive performance of existing clinical, radiographic, and combined scores

Predictive performance of established scores for outcome prediction after aSAH ranged between very low (AUC 0.55, original Fisher score) and moderately good (AUC 0.76, Hunt and Hess score and GCS score). The performance of the other scores showed similar ranges. The predictive performances of machine learning models were comparable with traditional GLM methods. For an overview of the performance values and the measures of spread, see Table 2. Detailed results for all additional performance measures are presented in the Supplementary Material (Tables 1–8).

Predictive performances of the combined set of clinico-radiological features

The combined set of clinical and radiographic features showed an AUC of 0.78 for the tree boosting model and 0.77 for all other models with the exception of the Naive Bayes model (0.75) (Table 3, Fig. 1A). There was no apparent superiority of the combined model over single clinical score models. The feature importance rating identified the GCS score as the most important feature in all models (Fig. 1B). Consistently, the second most important feature was age. The models also assigned importance to BNI and the presence of ICH. The Naive Bayes model was the only model assigning very high importance to pupil status. Results for the additional performance measures are presented in the supplementary material (Suppl. Tables).

Estimation of calibration

Based on the Brier score, the calibration was sufficient, ranging from 0.18 to 0.25 over all models. The best calibrated models were the combined set, the GCS model, and the Hunt and Hess score model (Table 4).

Discussion

In this study on aSAH outcome prediction, we observed moderately good performances of ML methods using traditional clinico-radiographic features available on admission. There was no difference in performance between any of the applied techniques, especially not between the traditional techniques (GLM), and the most modern techniques (CatBoost tree boosting, MLP). Furthermore, we observed no superiority of the examined ML techniques over the best performing clinical scores on admission (GCS and Hunt and Hess). Thus, we could not establish a relevant advantage of state-of-the-art ML methods for aSAH outcome prediction by using patient-specific clinical and radiographic features available on admission.

Outcome prediction in aSAH is usually conducted using traditional clinical and radiological scores on patient admission. Outcome prediction models find use in counseling of patients and their relatives as well as in the selection of treatment strategies. Especially in the presence of an ongoing global pandemic, precise predictions of outcomes in critically ill

Table 2 Clinical, radiographic, and treatment characteristics of patients with aSAH. Pathological pupil reaction describes a pupil reaction other than pupils equal and reactive to light. *WFNS* World Federation of Neurological Societies, *IVH* intraventricular hemorrhage, *ICH* intracerebral hemorrhage, *SDH* subdural hemorrhage, *SAH* subarachnoidal hemorrhage, *ACA* anterior cerebral artery, *MCA* middle cerebral artery, *ICA* internal cerebral artery, *mRS* Modified Rankin Scale. Localization of the aneurysm was available for 380/388 patients

			% (*n*)
Clinical features	Pathological pupil reaction		13.4% (52)
	GCS at admission	3	32.3% (125)
		4–8	8.7% (34)
		9–12	9.0% (35)
		13–15	50.0% (194)
Clinical scores	WFNS	I	36.6% (142)
		II	9.8% (38)
		III	3.4% (13)
		IV	12.6% (49)
		V	37.6% (146)
	Hunt and Hess	I	24.2% (94)
		II	17.5% (68)
		III	14.7% (57)
		IV	14.2% (55)
		V	29.4% (114)
Radiographic features	IVH		44.3% (172)
	ICH		32.0% (124)
	SDH		6.5% (25)
	Midline shift (> = 5 mm)		23.1% (89)
	Thickness of SAH (BNI)	< 5 mm (1°)	6.4% (25)
		6–10 mm (2°)	16.0% (62)
		11–15 mm (3°)	29.9% (116)
		15–20 mm (4°)	32.0% (124)
		> 25 mm (5°)	15.7% (61)
	Aneurysm location	ACA	35.8% (136)
		ICA	19.2% (73)
		MCA	26.1% (99)
		Posterior circulation	18.9% (72)
Radiographic scores	Modified Fisher	0	4.6% (18)
		1	12.1% (47)
		2	5.9% (23)
		3	26.8% (104)
		4	50.5% (196)
Combined score	VASOGRADE	Green	15.9% (62)
		Yellow	34.1% (132)
		Red	50.0% (194)
Outcome	Favorable	mRS 0	22.2% (86)
		mRS 1	18.3% (71)
		mRS 2	6.2% (24)
		mRS 3	7.2% (28)
	Unfavorable	mRS 4	7.7% (30)
		mRS 5	4.9% (19)
		mRS 6	33.5% (130)
Treatment	Coiling		57.9% (220)
	Clipping		27.9% (106)
	Other		2.6% (10)
	None		10.9% (25)

Table 3 Predictive performance of clinical, radiological, and combined scores as well as the combined feature set (see “Materials and methods” section) for unfavorable patient outcome (mRS 3–6) measured by AUC. Median AUC for the training and the test set (in bold) as well as the interquartile range (IQR) for the test set (in brackets) over 50 shuffles are shown. *AUC* area under the curve, *BNI* Barrow Neurological Institute scale, *GCS* Glasgow Coma Scale, *GLM* generalized linear model, *ICH* intracerebral hemorrhage, *IVH* intraventricular hemorrhage, *MLP* multilayer perceptron, *mRS* Modified Rankin Scale, *NB* Naive Bayes, *WFNS* World Federation of Neurological Societies, *SAH* subarachnoidal hemorrhage, *SDH* subdural hemorrhage, *SVMC* support vector machine classifier

Features	GLM	GLM_Lasso	GLM_elastic_net	CatBoost	MLP	SVMC	NB
Hunt and Hess score	0.75/0.76 (0.07)	0.75/0.75 (0.08)	0.75/0.0.75 (0.08)	0.75/0.76 (0.07)	0.75/0.76 (0.07)	0.75/0.75 (0.07)	0.75/0.76 (0.07)
WFNS score	0.74/0.74 (0.04)	0.74/0.74 (0.04)	0.73/0.74 (0.09)	0.74/0.74 (0.04)	0.74/0.74 (0.04)	0.74/0.74 (0.05)	0.74/0.74 (0.04)
Modified Fisher score	0.65/0.65 (0.07)	0.65/0.65 (0.07)	0.64/0.62 (0.15)	0.65/0.65 (0.07)	0.64/0.65 (0.07)	0.65/0.64 (0.07)	0.65/0.65 (0.07)
Original Fisher score	0.55/0.55 (0.04)	0.55/0.55 (0.05)	0.55/0.52 (0.11)	0.55/0.55 (0.06)	0.55/0.55 (0.04)	0.49/0.47 (0.11)	0.55/0.54 (0.08)
VASOGRADE score	0.72/0.72 (0.07)	0.72/0.72 (0.07)	0.72/0.72 (0.09)	0.72/0.71 (0.06)	0.72/0.72 (0.06)	0.72/0.72 (0.07)	0.72/0.72 (0.07)
BNI score	0.62/0.0.63 (0.06)	0.62/0.63 (0.06)	0.61/0.60 (0.15)	0.62/0.62 (0.07)	0.62/0.62 (0.07)	0.62/0.62 (0.08)	0.62/0.63 (0.06)
GCS score	0.75/0.76 (0.05)	0.75/0.76 (0.05)	0.75/0.75 (0.07)	0.76/0.76 (0.06)	0.75/0.76 (0.06)	0.75/0.76 (0.05)	0.75/0.76 (0.05)
Age, GCS score, sex, pupil status, presence of IVH, presence of ICH, presence of midline shift > 5 mm, presence of SDH, localization anterior circulation or other, BNI score	0.79/0.77 (0.06)	0.78/0.77 (0.06)	0.77/0.77 (0.07)	0.82/0.78 (0.07)	0.78/0.77 (0.06)	0.78/0.77 (0.06)	0.76/0.75 (0.07)

patients may help allocate scarce medical resources [4, 11, 33]. Therefore, the transparency, comparability, and reproducibility of outcome prediction models are of utmost importance. Recently, the comparability of clinical, radiographic, and combined scores in the same patient cohort was established. A combination of clinical and radiographic elements within single combined scores (VASOGRADE) did not show a significant improvement of predictive score performance regarding the prediction of angiographic vasospasm, cerebral infarction, and unfavorable outcome [9]. Another study showed that even patients with the highest Hunt and Hess score (V) have favorable outcome in 26 % of cases in a retrospective multicenter series [36].

The majority of previously established aSAH outcome prediction models are based on neurological deficits on admission and radiographic features, such as thickness of subarachnoid blood clots and the presence of IVH or ICH. However, more recent evidence suggests that other factors play a role in precise outcome prediction, such as patient age, pupil status, and aneurysm size and location[21, 29]. The inclusion of a high number of variables is one of the main strengths of ML approaches. In numerous medical fields, ML-based prediction models were shown to be superior to traditional techniques [12, 38]. In neurosurgery, ML prediction models have been evaluated for a variety of pathologies with variable predictive performances (AUC 0.71 to 0.96) [26]. In the prediction of the occurrence of shunt-dependent hydrocephalus after aSAH, ML methods proved to be superior to traditional methods [24]. They included dynamic variables such as infections, treatment timing from symptom onset, and fever onset. In predicting early complications after intracranial tumor surgery, ML methods showed slight superiority over conventional traditional methods [30]. In our present study, we applied—amongst others—two of the most promising state-of-the-art ML techniques to predict functional outcome after aSAH: tree boosting and ANN. Both have shown considerable advances over traditional linear or logistic regression techniques in the past [12, 38], even though traceability and comparability across different studies is reduced by substantial heterogeneity of clinical questions, input and output variables, and applied algorithms [19, 24, 26, 30].

To maintain transparency and comparability to existing models, our current approach uses established scoring systems. We applied a variety of ML techniques to the same dataset and acquired rather equivalent results regarding predictive power, sensitivity, and specificity but some difference regarding feature rating. Our analyses showed no superiority of any of the examined ML methods over traditional methods for aSAH outcome prediction. A combined set of relevant radiological and clinical features showed only a small superiority to simple and established clinical scores (e.g., tree boosting on the combined

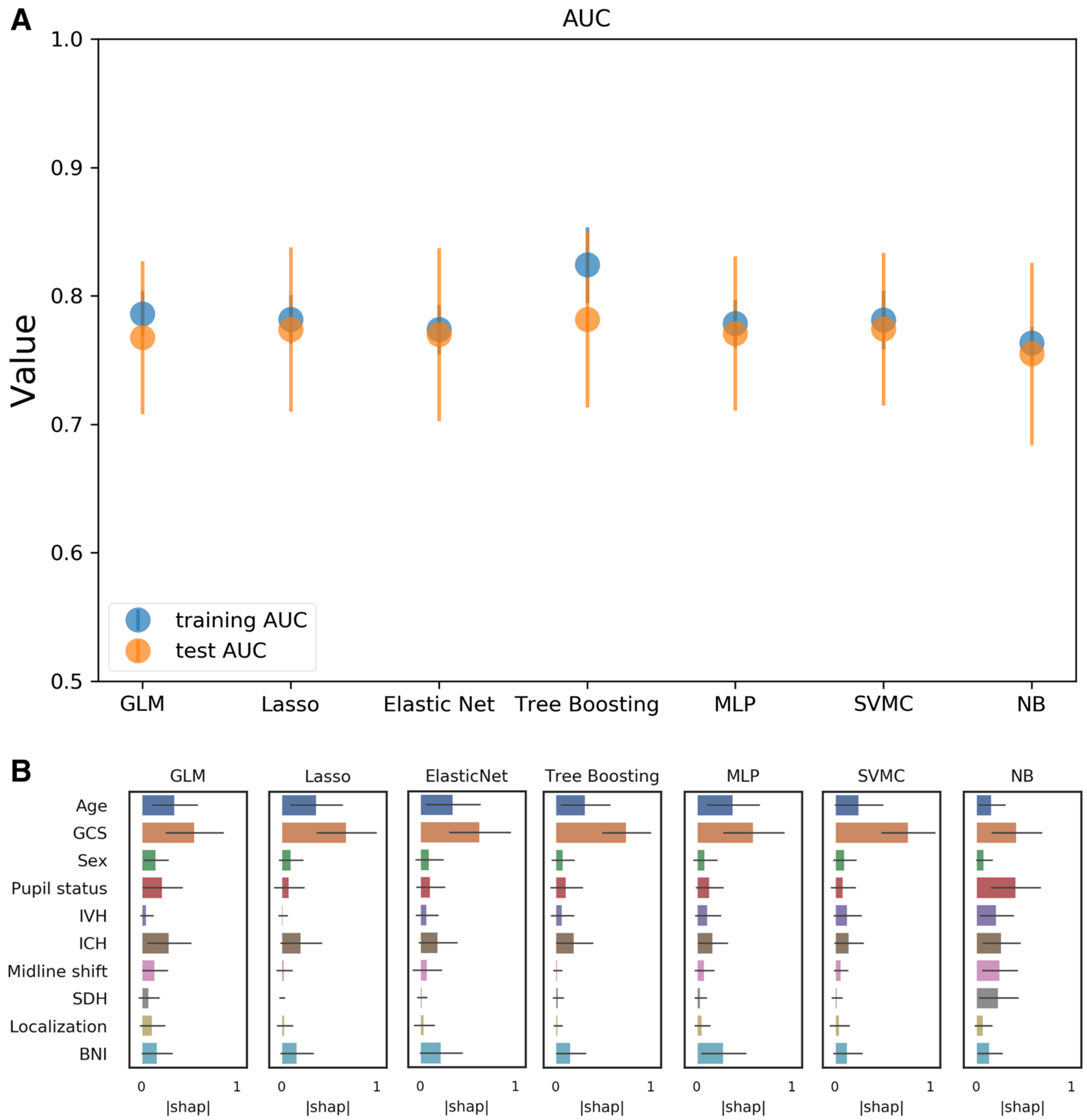

Fig. 1 Graphical representation of the performance and feature rating for the clinico-radiological model. **A** The highest test-AUC was 0.78 for the tree boosting model, with the exception of NB (0.75); the other models had a test-AUC value of 0.77. A larger difference between training and test set was observed for the tree boosting model indicative of overfitting. **B** The feature importance rankings consistently identified GCS as the most important factor. Note that model 7, GCS alone, already reached a test-AUC of 0.76. *AUC* area under the curve, *GCS* Glasgow Coma Scale, *GLM* generalized linear model, *IVH* presence of intraventricular hemorrhage, *ICH* presence of intracranial hemorrhage, *NB* Naive Bayes, *MLP* multilayer perceptron, *SDH* presence of subdural hematoma, *SVMC* support vector machine classifier, *BNI* semi-quantitative analysis of the thickness of subarachnoidal blood with respect to the scale introduced by the Barrow Neurological Institute in 2012[7]. The term "localization" refers to the localization of the aneurysm (anterior circulation yes/no)

features vs. Hunt and Hess and GCS alone). This was also shown for a decision tree model that reached similar accuracy than logistic regression in another study [7]. Notably, one of the main advantages of ML methods is their ability to capture even weak interactions between variables to make predictions. Nevertheless, our findings suggest that currently available scores and variables used to feed ML-based prediction models for aSAH may not contain enough information to improve the accuracy of outcome predictions.

Thus, it is warranted to explore the addition of other features available on patient admission in future works on early prediction models. These features could include laboratory data, imaging source data, and comorbidities. Also, events occurring during later phases of the course of aSAH, such as infectious diseases (e.g., pneumonia, meningitis), or

Table 4 Brier score results for clinical, radiological, and combined scores as well as the combined feature set (see "Materials and methods" section) for prediction of unfavorable patient outcome (mRS 3–6). Median Brier score for the training and the test set (in bold) as well as the interquartile range (IQR) for the test set (in brackets) over 50 shuffles are shown. *AUC* area under the curve, *BNI* Barrow Neurological Institute scale, *GCS* Glasgow Coma Scale, *GLM* generalized linear model, *ICH* intracerebral hemorrhage, *IVH* intraventricular hemorrhage, *MLP* multilayer perceptron, *mRS* Modified Rankin Scale, *NB* Naive Bayes, *WFNS* World Federation of Neurological Surgeons, *SAH* subarachnoidal hemorrhage, *SDH* subdural hemorrhage, *SVMC* support vector machine classifier

Features	GLM	GLM_Lasso	GLM_elastic_net	CatBoost	MLP	SVMC	NB
Hunt and Hess score	0.20/0.20 (0.02)	0.20/0.20 (0.02)	0.24/0.0.24 (0.06)	0.20/0.20 (0.02)	0.21/0.23 (0.05)	0.20/0.20 (0.02)	0.20/0.20 (0.03)
WFNS score	0.20/0.20 (0.02)	0.20/0.20 (0.01)	0.23/0.22 (0.06)	0.20/0.20 (0.02)	0.23/0.24 (0.05)	0.20/0.20 (0.02)	0.20/0.20 (0.02)
Modified Fisher score	0.23/0.23 (0.02)	0.24/0.24 (0.01)	0.25/0.25 (0.03)	0.23/0.23 (0.02)	0.24/0.25 (0.02)	0.23/0.23 (0.01)	0.24/0.24 (0.02)
Original Fisher score	0.25/0.25 (0.01)	0.25/0.25 (0.00)	0.28/0.27 (0.06)	0.25/0.25 (0.01)	0.25/0.25 (0.00)	0.25/0.25 (0.00)	0.25/0.25 (0.02)
VASOGRADE score	0.21/0.21 (0.02)	0.21/0.21 (0.02)	0.24/0.24 (0.07)	0.20/0.20 (0.03)	0.20/0.20 (0.03)	0.20/0.21 (0.03)	0.21/0.21 (0.03)
BNI score	0.24/0.0.24 (0.01)	0.24/0.24 (0.01)	0.25/0.25 (0.04)	0.24/0.24 (0.01)	0.25/0.25 (0.01)	0.24/0.24 (0.01)	0.24/0.24 (0.01)
GCS score	0.20/0.20 (0.02)	0.20/0.20 (0.02)	0.24/0.23 (0.05)	0.19/0.19 (0.03)	0.21/0.23 (0.05)	0.20/0.20 (0.02)	0.20/0.20 (0.03)
Age, GCS score, sex, pupil status, presence of IVH, presence of ICH, presence of midline shift > 5 mm, presence of SDH, localization anterior circulation or other, BNI score	0.19/0.19 (0.03)	0.19/0.19 (0.03)	0.20/0.21 (0.03)	0.18/0.19 (0.03)	0.19/0.20 (0.04)	0.19/0.19 (0.03)	0.24/0.23 (0.07)

cardiovascular complications (e.g., Takotsubo myocarditis) may be added over time to improve predictive performances [5, 16, 20]. General scores with special focus on physiology parameters shown to predict the course of intensive care treatment like the APACHE or SOFA scores could be added as well. To our knowledge, only one other work used ML techniques to predict outcome after aSAH [19]. While the analysis was performed in a large prospective multicenter cohort of aSAH patients, in that work outdated methodology, selection of features beyond admission, the lack of reported AUC, and Glasgow Outcome Score as the outcome measure make the models clinically less applicable and not comparable to our work.

Limitations of our study

Limitations of our study include the retrospective, single-center study design impacting the availability of features. Our patient sample is medium sized compared to existing studies applying ML methods for aSAH outcome prediction [7, 19]. However, a selection bias applies for most other studies that analyze aSAH as they are often taken from multicenter trial data with specific study protocols and inclusion/exclusion criteria. Our data represent real-world data from a single high-volume center in Germany. Our results may therefore not be generalized to other centers or countries [10]. Mean/mode imputation is not a state-of-the-art imputation method. State-of-the-art imputation methods are currently not tailored to predictive modelling, i.e., the transfer of imputation models from training to test set is not straightforward. However, given the very low ratio of missing values in our study, we deem this issue negligible and encourage the development of methods allowing the transfer of imputation models tailored to predictive modelling in Python. The very small imbalance in dichotomized outcome numbers may cause negligible bias. It is thus acknowledged but did not warrant a sub-sampling approach limiting the available data for model training.

Conclusion

Our study applies ML techniques for functional outcome prediction after aSAH on the basis of clinico-radiographic variables available at patient admission. We could demonstrate that the predictive performance of ML techniques was comparable but not superior to established traditional methods and established clinical scores. In conclusion, our findings make a compelling case for the exploration of new input variables other than traditional clinico-radiographic features to achieve a higher accuracy for outcome prediction in aSAH in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10143-020-01453-6.

Availability of data and material Data is available on demand.

Author contributions NFD: conception and design of the study, data analysis, draft and revision of the manuscript

VIM: data analysis and interpretation, draft and revision of the manuscript, creation of new software

MU: data analysis

EZ: administrative, technical and material support (creation of new software)

SCB: data analysis

AH: administrative, technical and material support (creation of new software)

ML: administrative, technical and material support (creation of new software)

SW: data analysis and interpretation, administrative, technical and material support

PV: administrative, technical and material support

DF: study supervision, drafting and the article, revision of manuscript

All authors read, critically revised and approved the final version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has received funding by 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 for the research group PREDICTioN2020 (lead: DF).

Compliance with ethical standards

Disclosures ND was funded by the institutional Rahel Hirsch and Lydia Rabinowitsch scholarships, received public body funding for the project GoSafe (Horizon 2020), and accepted speaker honoraria of Integra LifeSciences. No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethical approval The study was approved by the ethics review board of Charité Universitätsmedizin Berlin (EA1/291/14).

Informed consent for participation and publication Due to the retrospective nature of the study, no informed consent was obtained in accordance with our ethics review board.

Previous presentations A modified version of the abstract is going to be presented at the 71th annual meeting of the German Society for Neurosurgery (DGNC) on June 24, 2020 (originally planned to take place in Lübeck, Germany. Now set as digital conference).

Code availability https://github.com/prediction2020/explainable-predictive-models

References

1. Brimblecombe FS, Stoneman ME (1969) Score for respiratory-distress syndrome. Lancet 1:946
2. Connolly ES Jr, Rabinstein AA, Carhuapoma JR, Derdeyn CP, Dion J, Higashida RT, Hoh BL, Kirkness CJ, Naidech AM, Ogilvy CS, Patel AB, Thompson BG, Vespa P, American Heart Association Stroke C, Council on Cardiovascular R, Intervention, Council on Cardiovascular N, Council on Cardiovascular S, Anesthesia, Council on Clinical C (2012) Guidelines for the management of aneurysmal subarachnoid hemorrhage: a guideline for healthcare professionals from the American Heart Association/american Stroke Association. Stroke 43:1711–1737
3. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, Manceron P, Toldo S, Zaslavskiy M, Le Stang N, Girard N, Elemento O, Nicholson AG, Blay JY, Galateau-Salle F, Wainrib G, Clozel T (2019) Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med 25:1519–1525
4. Dafer RM, Osteraas ND, Biller J (2020) Acute stroke care in the coronavirus disease 2019 pandemic. J Stroke Cerebrovasc Dis 29: 104881
5. Dasenbrock HH, Rudy RF, Smith TR, Guttieres D, Frerichs KU, Gormley WB, Aziz-Sultan MA, Du R (2016) Hospital-acquired infections after aneurysmal subarachnoid hemorrhage: a nationwide analysis. World Neurosurg 88:459–474
6. de Oliveira Manoel AL, Jaja BN, Germans MR, Yan H, Qian W, Kouzmina E, Marotta TR, Turkel-Parrella D, Schweizer TA, Macdonald RL, collaborators S (2015) The VASOGRADE: a simple grading scale for prediction of delayed cerebral ischemia after subarachnoid hemorrhage. Stroke 46:1826–1831
7. de Toledo P, Rios PM, Ledezma A, Sanchis A, Alen JF, Lagares A (2009) Predicting the outcome of patients with subarachnoid hemorrhage using machine learning techniques. IEEE Trans Inf Technol Biomed 13:794–801
8. Dengler NF, Diesing D, Sarrafzadeh A, Wolf S, Vajkoczy P (2017) The Barrow Neurological Institute scale revisited: predictive capabilities for cerebral infarction and clinical outcome in patients with aneurysmal subarachnoid hemorrhage. Neurosurgery 81:341–349
9. Dengler NF, Sommerfeld J, Diesing D, Vajkoczy P, Wolf S (2018) Prediction of cerebral infarction and patient outcome in aneurysmal subarachnoid hemorrhage: comparison of new and established radiographic, clinical and combined scores. Eur J Neurol 25:111–119
10. Dijkland SA, Jaja BNR, van der Jagt M, Roozenbeek B, Vergouwen MDI, Suarez JI, Torner JC, Todd MM, van den Bergh WM, Saposnik G, Zumofen DW, Cusimano MD, Mayer SA, Lo BWY, Steyerberg EW, Dippel DWJ, Schweizer TA, Macdonald RL, Lingsma HF, Members of the SC (2019) Between-center and between-country differences in outcome after aneurysmal subarachnoid hemorrhage in the Subarachnoid Hemorrhage International Trialists (SAHIT) repository. J Neurosurg:1–9
11. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, Zhang C, Boyle C, Smith M, Phillips JP (2020) Fair allocation of scarce medical resources in the time of Covid-19. N Engl J Med 382:2049–2055
12. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. Nat Med 25:24–29
13. Fisher CM, Kistler JP, Davis JM (1980) Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computerized tomographic scanning. Neurosurgery 6:1–9
14. Frontera JA, Claassen J, Schmidt JM, Wartenberg KE, Temes R, Connolly ES Jr, MacDonald RL, Mayer SA (2006) Prediction of

symptomatic vasospasm after subarachnoid hemorrhage: the modified fisher scale. Neurosurgery 59:21–27 discussion 21-27

15. Hunt WE, Hess RM (1968) Surgical risk as related to time of intervention in the repair of intracranial aneurysms. J Neurosurg 28:14–20
16. Kahn JM, Caldwell EC, Deem S, Newell DW, Heckbert SR, Rubenfeld GD (2006) Acute lung injury in patients with subarachnoid hemorrhage: incidence, risk factors, and outcome. Crit Care Med 34:196–202
17. Lee VH, Ouyang B, John S, Conners JJ, Garg R, Bleck TP, Temes RE, Cutting S, Prabhakaran S (2014) Risk stratification for the in-hospital mortality in subarachnoid hemorrhage: the HAIR score. Neurocrit Care 21:14–19
18. 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
19. Lo BW, Macdonald RL, Baker A, Levine MA (2013) Clinical outcome prediction in aneurysmal subarachnoid hemorrhage using Bayesian neural networks with fuzzy logic inferences. Comput Math Methods Med 2013:904860
20. Lo BW, Fukuda H, Angle M, Teitelbaum J, Macdonald RL, Farrokhyar F, Thabane L, Levine MA (2016) Clinical outcome prediction in aneurysmal subarachnoid hemorrhage - alterations in brain-body interface. Surg Neurol Int 7:S527–S537
21. Mader MM, Piffko A, Dengler NF, Ricklefs FL, Duhrsen L, Schmidt NO, Regelsberger J, Westphal M, Wolf S, Czorlich P (2020) Initial pupil status is a strong predictor for in-hospital mortality after aneurysmal subarachnoid hemorrhage. Sci Rep 10:4764
22. Miles J (2014) Tolerance and Variance Inflation Factor. Wiley Stats:Ref: Statistics Reference Online (American Cancer Society, 2014)
23. Mullie A, Verstringe P, Buylaert W, Houbrechts H, Michem N, Delooz H, Verbruggen H, Van den Broeck L, Corne L, Lauwaert D et al (1988) Predictive value of Glasgow coma score for awakening after out-of-hospital cardiac arrest. Cerebral Resuscitation Study Group of the Belgian Society for Intensive Care. Lancet 1: 137–140
24. Muscas G, Matteuzzi T, Becattini E, Orlandini S, Battista F, Laiso A, Nappini S, Limbucci N, Renieri L, Carangelo BR, Mangiafico S, Della Puppa A (2020) Development of machine learning models to prognosticate chronic shunt-dependent hydrocephalus after aneurysmal subarachnoid hemorrhage. Acta Neurochir (Wien) 162: 3093–3105
25. Sandow N, Diesing D, Sarrafzadeh A, Vajkoczy P, Wolf S (2016) Nimodipine dose reductions in the treatment of patients with aneurysmal subarachnoid hemorrhage. Neurocrit Care 25:29–39
26. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, Smith TR, Arnaout O (2018) Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg 109:476–486.e471
27. Shailaja K, Seetharamulu B, Jabbar MA (2018) Machine learning in healthcare: a review. 2018 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA)
28. Teasdale GM, Drake CG, Hunt W, Kassell N, Sano K, Pertuiset B, De Villiers JC (1988) A universal subarachnoid hemorrhage scale: report of a committee of the World Federation of Neurosurgical Societies. J Neurol Neurosurg Psychiatry 51:1457
29. van Donkelaar CE, Bakker NA, Birks J, Veeger N, Metzemaekers JDM, Molyneux AJ, Groen RJM, van Dijk JMC (2019) Prediction of outcome after aneurysmal subarachnoid hemorrhage. Stroke 50: 837–844
30. van Niftrik CHB, van der Wouden F, Staartjes VE, Fierstra J, Stienen MN, Akeret K, Sebok M, Fedele T, Sarnthein J, Bozinov O, Krayenbuhl N, Regli L, Serra C (2019) Machine learning algorithm identifies patients at high risk for early complications after intracranial tumor surgery: registry-based cohort study. Neurosurgery 85:E756–E764
31. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J (1988) Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19:604–607
32. VI HDaM (2020) From bit to bedside: a practical framework for artificial intelligence product development in healthcare. Advanced intelligent systems 2
33. Vincent JL, Creteur J (2020) Ethical aspects of the COVID-19 crisis: How to deal with an overwhelming shortage of acute beds. Eur Heart J Acute Cardiovasc Care 9:248–252
34. Wartenberg KE, Hwang DY, Haeusler KG, Muehlschlegel S, Sakowitz OW, Madzar D, Hamer HM, Rabinstein AA, Greer DM, Hemphill JC 3rd, Meixensberger J, Varelas PN (2019) Gap analysis regarding prognostication in neurocritical care: a joint statement from the German Neurocritical Care Society and the Neurocritical Care Society. Neurocrit Care 31:231–244
35. Wilson DA, Nakaji P, Abla AA, Uschold TD, Fusco DJ, Oppenlander ME, Albuquerque FC, McDougall CG, Zabramski JM, Spetzler RF (2012) A simple and quantitative method to predict symptomatic vasospasm after subarachnoid hemorrhage based on computed tomography: beyond the Fisher scale. Neurosurgery 71:869–875
36. Wostrack M, Sandow N, Vajkoczy P, Schatlo B, Bijlenga P, Schaller K, Kehl V, Harmening K, Ringel F, Ryang YM, Friedrich B, Stoffel M, Meyer B (2013) Subarachnoid haemorrhage WFNS grade V: is maximal treatment worthwhile? Acta Neurochir (Wien) 155:579–586
37. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O, written on behalf of AMEB-DCTCG (2019) Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med 7:152
38. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, Frey D (2020) Opening the black box of artificial intelligence for clinical decision support: a study predicting stroke outcome. PLoS One 15:e0231166

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