Neurocrit Care (2024) 40:1171–1181
https://doi.org/10.1007/s12028-023-01832-z

NEUROCRITICAL CARE SOCIETY

REVIEW ARTICLE

Machine Learning Algorithms to Predict Delayed Cerebral Ischemia After Subarachnoid Hemorrhage: A Systematic Review and Meta-analysis

Laís Silva Santana[1], Jordana Borges Camargo Diniz[2], Nicollas Nunes Rabelo[3], Manoel Jacobsen Teixeira[3], Eberval Gadelha Figueiredo[3] and João Paulo Mota Telles[2*]

Abstract

Delayed cerebral ischemia (DCI) is a common and severe complication after subarachnoid hemorrhage (SAH). Logistic regression (LR) is the primary method to predict DCI, but it has low accuracy. This study assessed whether other machine learning (ML) models can predict DCI after SAH more accurately than conventional LR. PubMed, Embase, and Web of Science were systematically searched for studies directly comparing LR and other ML algorithms to forecast DCI in patients with SAH. Our main outcome was the accuracy measurement, represented by sensitivity, specificity, and area under the receiver operating characteristic. In the six studies included, comprising 1828 patients, about 28% (519) developed DCI. For LR models, the pooled sensitivity was 0.71 (95% confidence interval [CI] 0.57–0.84; $p < 0.01$) and the pooled specificity was 0.63 (95% CI 0.42–0.85; $p < 0.01$). For ML models, the pooled sensitivity was 0.74 (95% CI 0.61–0.86; $p < 0.01$) and the pooled specificity was 0.78 (95% CI 0.71–0.86; $p = 0.02$). Our results suggest that ML algorithms performed better than conventional LR at predicting DCI.

Trial Registration: PROSPERO (International Prospective Register of Systematic Reviews) CRD42023441586; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=441586

Keywords: Machine learning, Delayed cerebral ischemia, Logistic regression, Subarachnoid hemorrhage, Prediction model

Introduction

Delayed cerebral ischemia (DCI) is one of the most frequent complications after subarachnoid hemorrhage (SAH) and is sometimes a determinant of poor prognosis due to late diagnosis [1, 2]. The early identification of DCI can either interfere with the patient's prognosis or reduce the costs of intensive care.

Logistic regression (LR) is the conventional method for predicting DCI, but it has limitations including low precision and complexity in the use of data, given the multicollinearity between the variables [3]. For this reason, machine learning (ML) algorithms appear as a potential alternative for the analysis of clinical data, because in addition to being able to process large amounts of data, they can learn the parameters, optimizing the obtained results [4].

Some studies involving DCI prediction through clinical data suggest that ML models have greater predictive power than LR [5–7]. Other studies that have applied ML algorithms using heterogeneous data (including both clinical information and imaging tests) have shown

*Correspondence: joao.telles@fm.usp.br
[2] Department of Neurology, University of São Paulo, São Paulo, SP, Brazil
Full list of author information is available at the end of the article

Springer

positive results in both the diagnosis of Alzheimer's disease [8] and the prediction of the risk of aortic stenosis [9].

Therefore, we aimed to perform a systematic review and meta-analysis comparing the effectiveness of ML models versus LR models to predict DCI after SAH, specifically interested in parameters such as sensitivity, specificity, and area under the receiver operating characteristic (AUROC).

Methods

Protocol and Registration

This systematic review and meta-analysis were reported based on recommendations from the Cochrane Collaboration and the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement guidelines. The review protocol was registered with PROSPERO (CRD42022383937).

Eligibility Criteria

We screened studies by title, abstract, and full text using the eligibility criteria defined by the elements of our question: population, intervention, control, and outcomes. The essential items consisted of the following:

1. *Population* men or women with SAH.
2. *Intervention* classic ML algorithms, such as random forest, support vector machine, and artificial neural networks.
3. *Control* conventional LR.
4. *Outcomes* accuracy measurement (sensitivity, specificity, AUROC, and accuracy).

We only selected studies that reported outcome measures related to the effectiveness of DCI predictive models, such as sensitivity, specificity, accuracy, confusion matrix, and so on. We excluded the following types of articles: reviews, case reports, editorials, correspondences, studies without peer review, and abstracts.

Search Strategy and Data Extraction

We searched PubMed, Embase, and Web of Science up to December 2022. No publication period limits were applied. The following search terms were included: ("subarachnoid hemorrhage" or "SAH" or "subarachnoid hemorrhages") AND ("delayed cerebral ischemia" or "DCI") AND (("machine learning" or "ML") or ("logistic regression" or "LR")). This search strategy was applied to all databases.

In addition to the main outcome measurements, the following baseline characteristics were collected: (1) number of patients, (2) sex distribution, (3) mean age, (4) hypertension, (5) proportion of patients with diabetes, (6) Hunt and Hess grade, and (7) modified Fisher scale. These data were extracted by two authors independently following predefined search criteria and quality assessment.

End Point

The main outcome was the accuracy measurement (sensitivity, specificity, AUROC, and accuracy) of the DCI prediction models.

Quality Assessment and Risk of Bias

This article used the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) criteria to evaluate the risk of bias. Domains included patient selection, index test, reference standard, and flow and timing. In line with the recommendations from the QUADAS-2 guidelines, questions per domain were tailored for this article and can be found in "Supplemental Materials, Adapted QUADAS-2 questions." If one of the questions was scored at risk of bias, the domain was scored as high risk of bias. At least one domain at high risk of bias resulted in an overall score of high risk of bias, and only one domain scored as unclear risk of bias resulted in an overall score of unclear risk of bias for that article. Quality assessment was performed by two investigators (JBCD, LSS) independently. Disagreements between the two investigators were solved through a consensus after a discussion among the authors and senior author.

Summary measures

Because of the diversity of data reported by the models under study, we chose to compare the AUROCs of each article. When the study did not report the AUROC, we estimated the metric using sensitivity and specificity (Eq. 1). Only one study [15] provided information on the time point at which the AUROCs were calculated, which is crucial for assessing the time-sensitivity of the features. In that study, the time point with the highest performance was chosen, specifically 7 days before the onset of DCI.

$$\begin{aligned} \text{AUROC} = {} & 0.5 \times \left(1 - \text{specificity}\right) \times \text{sensitivity} \\ & + \text{specificity} \times \text{sensitivity} + 0.5 \\ & \times \left(1 - \text{sensitivity}\right) \times \text{specificity} \end{aligned} \tag{1}$$

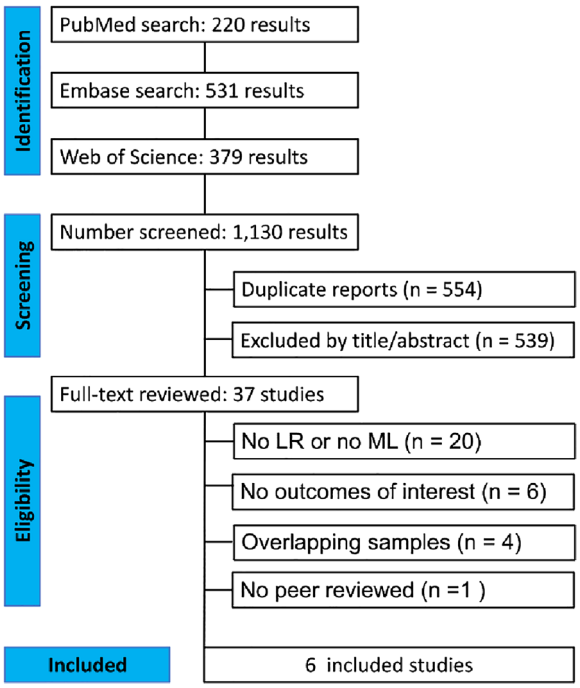

Fig. 1 PRISMA flow diagram of study screening and selection. *LR* logistic regression, *ML* machine learning, *PRISMA* preferred reporting items for systematic reviews and meta-analyses

Statistical Analysis

We extracted information on the true positives, true negatives, false positives, and false negatives and entered the data into Review Manager 5.4.1 (Nordic Cochrane Center, The Cochrane Collaboration, Copenhagen, Denmark) to calculate pooled measures of sensitivity and specificity, as well as the corresponding 95% confidence intervals (CIs). We also used R (R Foundation for Statistical Computing, Vienna, Austria, 2021) to perform a meta-analysis of the tests' performance.

Results

As illustrated by Fig. 1, a total of 1130 records were initially retrieved from three literature databases. Following the removal of duplicated records and the exclusion of studies that were not related to the topic of this meta-analysis, 37 studies remained for the full review. From those, six studies were deemed eligible.

In this meta-analysis, we enlisted a nonoverlapping sample of 1828 participants, of whom 519 (28%) developed DCI. The studies reported 627 male patients and 1255 female patients, with a mean age ranging from 53 to 55 years. More details regarding the characteristics of the eligible studies are presented in Table 1. Of the 29 models identified in this review, the majority of employment was represented by LR (11 models) and random forest (RF) (5 models), as represented in Table 2.

Regarding DCI diagnostic criteria, the majority of studies relied on a combination of clinical and radiographic findings, including alterations in the Glasgow Coma Scale (GCS) score, the emergence of new focal neurological deficits, and the presence of ischemic infarcts on computed tomography (CT) or magnetic resonance imaging scans. In the context of prediction, a variety of covariates or predictors were employed, typically encompassing age, sex, clinical grades (Hunt–Hess and World Federation of Neurological Surgeons scale), treatment of aneurysms, and laboratory test results (hemoglobin, sodium, white blood cell count, platelet count, and creatinine levels). Furthermore, certain studies integrated specific features associated with subarachnoid hemorrhage, such as CT values and the presence of cerebral edema. Table 3 displays the covariates and diagnostic criteria for DCI used in the models.

The ML models exhibited better overall performance than the LR models, as evidenced by the pooled sensitivity and specificity, based on the data of five studies. One study was excluded of this analysis as it provided only the AUROC and lacked data of the confusion matrix [15]. For the LR models, sensitivity and specificity values were corresponding to 0.71 (95% CI 0.57–0.84; Fig. 2) and 0.63 (95% CI 0.42–0.85; Fig. 3), respectively, whereas the equivalent values for the ML models were 0.74 (95% CI 0.61–0.86; Fig. 4) and 0.78 (95% CI 0.71–0.86; Fig. 5). As depicted in Tables 4 and 5, ML models showed higher AUROC values than those obtained by LR models. This superiority is also supported by the scatter plot of AUROC values grouped by models (Fig. 6), and the forest plots (Supplement Figs. 3 and 4). Further comparisons between the algorithms are provided in Tables 4 and 5.

For a more specific evaluation, we conducted subanalyses of individual algorithms that presented a satisfactory amount of data (Supplemental Figs. 5–14). The extreme gradient boosting (XGBoost) model achieved the highest pooled sensitivity (0.89; 95% CI 0.80–0.89; Supplemental Fig. 13), whereas the artificial neural network (ANN) model had the highest specificity (0.81; 95% CI 0.71–0.92; Supplemental Fig. 6). In other subanalyses, only retrospective studies were included to reduce heterogeneity (Supplemental Figs. 15–18). Again, the ML models demonstrated better results than the LR models in terms of sensitivity (0.74; 95% CI 0.64–0.85; Supplemental Fig. 17) and specificity (0.77; 95% CI 0.72–0.82; Supplemental Fig. 18). The LR models had lower sensitivity (0.68; 95% CI 0.60–0.77; Supplemental Fig. 15) and specificity (0.51; 95% CI 0.27–0.75; Supplemental Fig. 16) when compared with the ML models.

Table 1 Study characteristics

Characteristic	Alexopoulos	Chen	Hu	Megjhani	Ramos	Savarraj
Date of enrollment	2013–2020	2011–2015	2019–2021	2006–2020	2011–2015	2009–2016
Patients	213	107	404	388	317	399
DCI	42	56	112	124	97	88
Sex (male/female)	56/157	32/75	157/247	115/273	106/211	161/290
Mean age	55	56	57	55	57	53
Hypertension	N/A	N/A	192	189	104	211
Diabetes	N/A	N/A	12	N/A	21	40
Hunt and Hess scale						
I–II	92	N/A	270	N/A	151	N/A
III	64		86		56	
IV	25		27		24	
V	30		21		79	
mFS (modified Fisher Scale) scale						
1–2	84	N/A	113	195	19	N/A
3	118		98	37	69	224
4	3		295	41	216	N/A
WFNS grade						
I–II	N/A	N/A	295	N/A	173	N/A
III			50		9	
IV			34	160	32	
V			25		24	
Timing of prediction	Between days 4 and 12 following the aneurysm rupture	At day 3	Between 4 and 30 days after the initial onset of SAH	Day 1 (anchor)	At admission	3rd day postadmission

DCI, delayed cerebral ischemia, mFS, xxx, N/A, not applicable, SAH, subarachnoid hemorrhage, WFNS, World Federation of Neurological Surgeons

Table 2 Models

Study	LR model	ML model (n)
Alexopoulos [12]	Conventional LR	RF (1), XGBoost (2), and ANN (1)
Chen [13]	Conventional LR	GBTM (2)
Hu [14]	Conventional LR	KNN (1), SVM (1), DT (1), RF (1), XGBoost (1), and ANN (1)
Megjhani [15]	LR (L2-regularized logistic regression)	RF(1), SVM-K (1), SVM-L (1), and EC (1)
Ramos [5]	Conventional LR	SVM (2), RFC (2), and MCLP (2)
Savarraj [7]	Conventional LR	ANN (1)

ANN artificial neural network, *AUROC* area under the receiver operating characteristic, *DT* decision tree model, *EA* epileptiform abnormality, *EC* ensemble classifier, *GBTM* group-based trajectory modeling, *HH* Hunt–Hess score, *KNN* K-nearest neighbor, *LASSO* least absolute shrinkage and selection operator, *LR* logistic regression, *MCA* middle cerebral artery, *MCLP* multilayer perceptron, *RF or RFC* random forest, *SK* support vector machine kernel, *SL* support vector machine linear, *SVM* support vector machine, *XGBoost* extreme gradient boosting

Comparative Analysis: Strengths and Weaknesses

Regarding the strengths and weaknesses of the ML models employed, a detailed comparison of each algorithm's characteristics is provided in Table 6. ANN demonstrates exceptional capabilities in capturing intricate patterns and nonlinear relationships. However, their utilization can be computationally demanding, and their inherent black-box nature makes interpretation challenging [16]. Decision tree models offer simplicity and robustness against outliers but are susceptible to overfitting and struggle with high-dimensional datasets. Ensemble classifiers improve prediction accuracy by combining multiple models, yet they require careful configuration and may lack interpretability [17]. Group-based trajectory modeling identifies distinct subgroups and provides valuable insights into population dynamics, albeit relying on predetermined trajectory groups [18]. K-nearest neighbor presents an intuitive approach for capturing complex relationships, yet computational demands can increase with larger datasets [19]. Multilayer perceptron exhibits nonlinear modeling capabilities, albeit at the cost of interpretability [20]. RF effectively handles nonlinear relationships and missing values, yet interpretation can be challenging [21]. Support vector machine excels at capturing nonlinear relationships but necessitate careful parameter tuning [22]. Lastly, XGBoost showcases proficiency in capturing complex patterns but requires

Table 3 Diagnostic criteria

Study	DCI diagnostic criteria	Covariants in model
Alexopoulos [12]	Worsening of Modified Glasgow coma score for 2 or more points, followed by CT or brain MRI confirmation demonstrating new ischemic infarcts since their admission	Demographic, clinical, and radiographic characteristics
Chen [13]	(1) No other etiology could have caused a permanent or temporary focal neurological impairment (such as aphasia, apraxia, hemianopia, or neglect) between 4 and 14 days after aSAH; and/or (2) the GCS score decreased by at least (2) Two points [either on one of its components (eye opening, verbal response, motor response), or on total score]; and/or (3) head CT scans revealed a low-density area that was not noticeable on admission or immediately after the operation, and there were no other causes except vasospasms between 4 and 30 days after aSAH	Clinical variables (HH, aneurysm treatment, MCA PSV and the presence or absence of EA, defined as seizures, epileptiform discharges, and rhythmic/periodic activity)
Hu [14]	(1) No other cause could have led to the occurrence of a permanent or temporary focal neurological impairment (such as aphasia, apraxia, hemianopia, or neglect) between 4 and 14 days after aSAH; (2) the GCS score decreased by at least two points (either on one of its components [eye opening, verbal response, and motor response] or on total score); and (3) head CT scans revealed a low-density area that was not noticeable on admission or immediately after the operation, and there were no other causes except vasospasms between 4 and 30 days after aSAH. (consistent with Vergouwen et al.)	CT value of subarachnoid hemorrhage, WBC count, neutrophil count, CT value of cerebral edema and monocyte count
Megjhani [15]	Delayed neurological deterioration defined as a $\geq$ 2-point change in GCS or new focal neurological deficit lasting for > 1 h and not associated with surgical treatment or a new cerebral infarct on brain imaging that is not attributable to any other causes	Demographics included in the model were age, sex, mFS, World Federation of Neurological Surgeons scale, HH grade, and GCS at NICU admission
Ramos [5]	Occurrence of new focal neurological impairment or a decrease of two points or more on the GCS (with or without new hypodensity on CT) that could not be attributed to other causes. (according to Vergouwen et al.)	Clinical variables (modified Fisher scale on admission, number, location, height, and width of aneurysm were determined based on CT angiography image data)
Savarraj [7]	Definition according to Vergouwen et al.[a]	Age and laboratory test results such as hemoglobin, sodium, WBC, platelets, and creatinine

CT computed tomography, *DCI* delayed cerebral ischemia, *EA* epileptiform abnormality, *GCS* Glasgow Coma Scale, *HH* Hunt–Hess, *MCA* middle cerebral artery, *MR* magnetic resonance, *MRI* magnetic resonance imaging, *ICU* intensive care unit, *PSV* peak systolic velocity, *WBC* white blood cells

[a] Vergouwen et al.: "The presence of cerebral infarction on CT or MR scan of the brain within 6 weeks after SAH, or on the latest CT or MR scan made before death within 6 weeks, or proven at autopsy, not present on the CT or MR scan between 24 and 48 h after early aneurysm occlusion, and not attributable to other causes such as surgical clipping or endovascular treatment. Hypodensities on CT imaging resulting from ventricular catheter or intraparenchymal hematoma should not be regarded as cerebral infarctions from DCI."

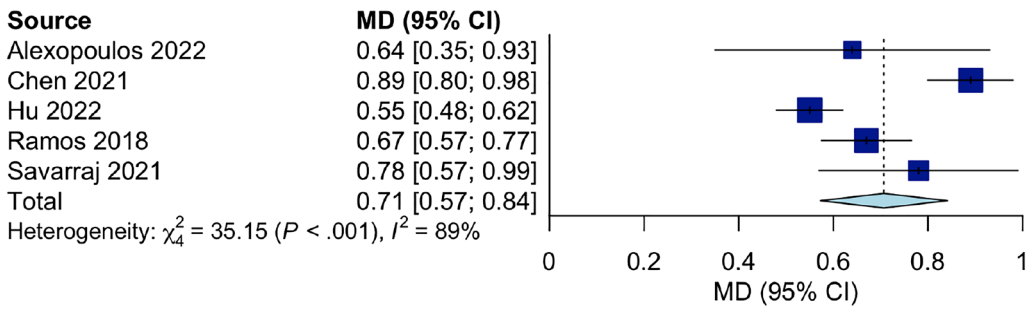

Fig. 2 Forest plot of sensitivity for logistic regression models. *CI* confidence interval, *MD* mean difference

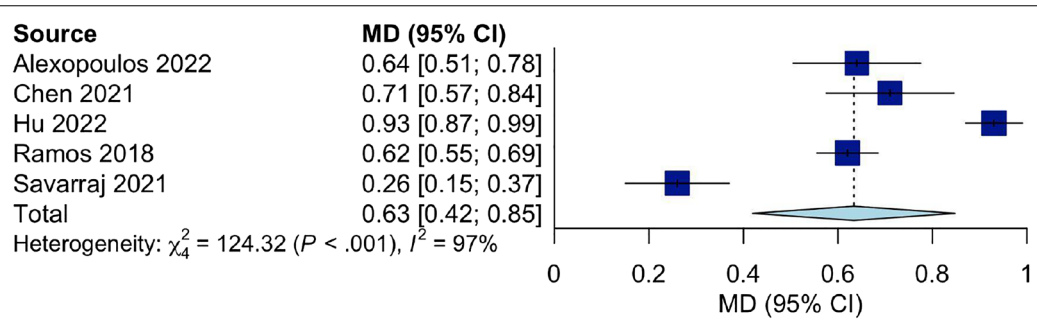

Fig. 3 Forest plot of specificity for logistic regression models. *CI* confidence interval, *MD* mean difference

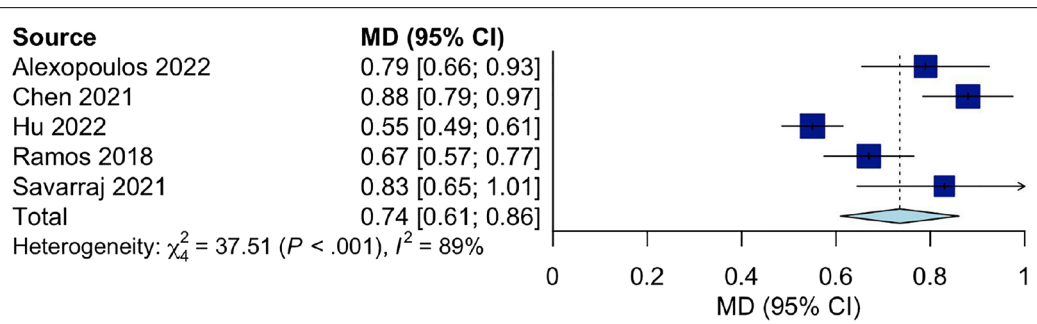

Fig. 4 Forest plot of sensitivity for machine learning models. *CI* confidence interval, *MD* mean difference

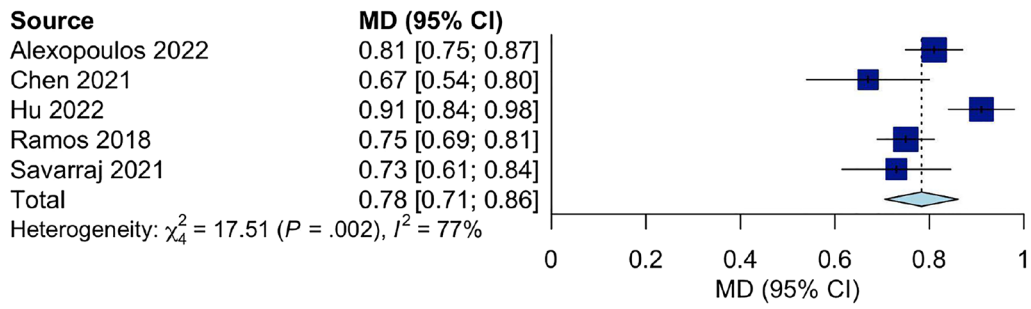

Fig. 5 Forest plot of specificity for machine learning models. *CI* confidence interval, *MD* mean difference

Table 4 Best performance of each study

Study	AUROC		Best performance
	LR	ML	
Alexopoulos [12]	0.653	0.866	ML
Chen [13]	0.7733	0.7517	LR
Hu [14]	0.824	0.858	ML
Megjhani [15]	0.68	0.89	ML
Ramos [5]	0.65	0.74	ML
Savarraj [7]	0.55	0.75	ML

AUROC area under the receiver operating characteristic, *LR* logistic regression, *ML* machine learning

meticulous parameter tuning and may have limited interpretability [23].

Key Variables in High-Performing ML Models

One study [12] employed the XGBoost algorithm with Boruta feature selection and aneurysm type as predictors. The inclusion of aneurysm type as a predictor is of utmost importance due to its provision of valuable information about anatomical characteristics and the risk of rupture. The XGBoost algorithm is renowned for its capability to capture intricate interactions and nonlinear relationships among variables, which likely contributed to its enhanced performance in this study.

Two studies [7, 14] achieved the best performance by employing ANN models. In the study by Hu et al. [14], significant predictors were CT value of subarachnoid hemorrhage, white blood cell count, neutrophil count, CT value of cerebral edema, and monocyte count. Similarly, Savarraj et al. [7] attained the highest results by incorporating a combination of electronic medical record variables, such as age, hemoglobin, sodium, white blood cell count, platelets, creatinine, and the clinician-derived Hunt–Hess score, in their ANN model.

LR model revealed that high-middle cerebral artery velocity collected on day 3 after SAH was a more influential predictor for DCI compared with epileptiform abnormalities (day 3) [23]. However, the Hunt–Hess score alone did not perform as well as other features. Among the ML models, the multitrajectory feature (day 3) outperformed the epileptiform abnormalities trajectory feature (day 3). These findings highlight the importance of high-middle cerebral artery velocity and multitrajectory in DCI prediction, suggesting their superiority over Hunt–Hess score and epileptiform abnormalities trajectory features.

One study [15] observed promising results in the application of RF and ensemble classifier. The model's training incorporated standard vital sign measurements (heart rate, blood pressure, respiratory rate, and oxygen saturation) alongside routine demographic data collected for clinical purposes, such as age, sex, modified Fisher score, World Federation of Neurological Surgeons scale, Hunt–Hess grade, and GCS at Neurological Intensive Care Unit Admission.

Finally, one study [5] combined clinical variables with extracted image features using RFc (random forest classifier). The most relevant features, in order of importance, included image features, total brain volume, presence of intraparenchymal blood, time from ictus to CT, age, aneurysm height, presence of subdural blood, aneurysm width, and GCS.

Timing Impact on DCI Prediction Model Performance

With respect to the relationship between the timing of DCI prediction and the performance of models, five studies did not provided information on the relationship between the timing of DCI prediction and the performance of the DCI prediction models [5, 7, 12, 14, 34]. However, one study reported that the model using data from the DCI to 12 h before the onset of DCI had the best performance [15].

Quality Assessment and Risk of Bias

Regarding quality assessment and risk of bias, the results according to QUADAS-2 guidelines are shown in "Supplemental Table 1." Only one study [14] had a low overall risk of bias. Three [5, 12, 15] out of six articles received an unclear risk of bias score for not clarifying whether the index test results were interpreted without knowledge of the results of the reference standard. Two studies [7, 13] had a high risk of bias because they failed to describe their study population (patient selection) or had inappropriate exclusions. However, apart from that, the majority of the four domains scored low risk of bias in all studies.

Discussion

We conducted a prospectively registered systematic review and meta-analysis of literature comparing LR and ML algorithms for predicting DCI with SAH in a cohort of 1828 patients. Our results suggest that ML models show promise for outperforming LR models. ML models exhibited slightly higher pooled sensitivity and specificity, which indicate that they may be more effective in identifying at-risk patients. Additionally, some ML models achieved substantially higher AUROC values, implying greater overall accuracy.

Table 5 Performance ranking of all models

Study	AUROC	Rank
Alexopoulos [12]		
Extreme Gradient Boosting (Boruta selection + Aneurysm type)	0.866	1
Extreme Gradient Boosting (Entire dataset)	0.795	2
Random Forests (Boruta selection)	0.713	3
Lasso regression (Boruta selection)	0.670	4
Logistic regression (LASSO residuals)	0.669	5
Lasso regression (Entire dataset)	0.658	6
Logistic regression (Boruta selection)	0.653	7
Artificial neural network (Entire dataset)	0.455	8
Chen [13]		
LR—High-MCA velocity (Day 3) + EA (Day 3) + HH + Aneurysm Treatment	0.773	1
LR—EA (Day 3) + HH + Aneurysm Treatment	0.758	2
GBTM Multitrajectory (Day 3) + HH + Aneurysm Treatment	0.752	3
GBTM EA Trajectory (Day 3) + HH + Aneurysm Treatment	0.747	4
LR—High-MCA velocity (Day 3) + EA (Day 3) + HH	0.710	5
LR—High-MCA velocity (Day 3) + HH + Aneurysm Treatment	0.691	6
LR—HH + Aneurysm Treatment	0.674	7
Hunt–Hess	0.568	8
Hu [14]		
ANN	0.858	1
LR	0.824	2
KNN	0.792	3
XGBoost	0.780	4
SVM	0.677	5
DT	0.675	6
Megjhani [15]		
RF	0.890	1
EC	0.830	2
SL	0.740	3
SK	0.720	4
LR	0.680	5
Ramos [5]		
RFC—All clinical variables see combined with extracted image features	0.740	1
RFC—All clinical variables	0.680	2
SVM—All clinical variables see combined with extracted image features	0.680	2
MLP—All clinical variables see combined with extracted image features	0.670	3
LR—All clinical variables see combined with extracted image features	0.650	4
SVM—All clinical variables	0.640	5
LR—Model 1 (Prior knowledge variables)	0.630	6
MLP—All clinical variables	0.630	6
LR—All clinical variables	0.610	7
LR—Model 2 (Prior knowledge variables)	0.590	2
Savarraj [7]		
ML model (ANN)	0.750	1
Standard model (LR)	0.550	2

ANN artificial neural network, *AUROC* area under the receiver operating characteristic, *DT* decision tree model, *EA* epileptiform abnormality, *EC* ensemble classifier, *GBTM* group-based trajectory modeling, *HH* Hunt–Hess score, *KNN* K-nearest neighbor, *LASSO* least absolute shrinkage and selection operator, *LR* logistic regression, *MCA* middle cerebral artery, *MCLP* multilayer perceptron, *RF or RFC* random forest, *SK* support vector machine kernel, *SL* support vector machine linear, *SVM* support vector machine, *XGBoost* extreme gradient boosting

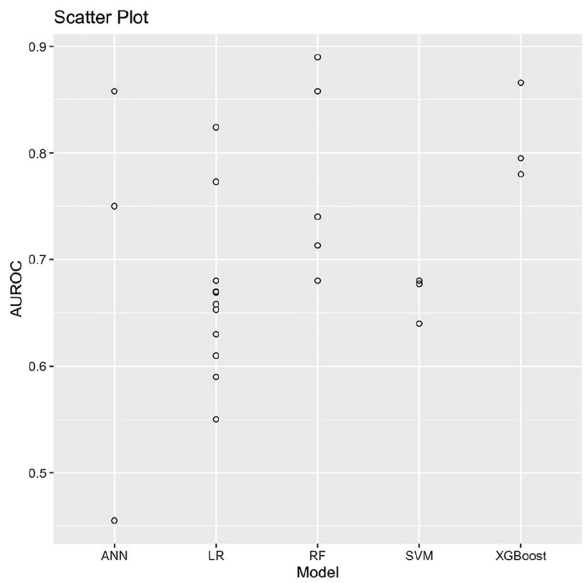

Fig. 6 Scatter plot of model performance. *ANN* artificial neural network, *AUROC* area under the receiver operating characteristic, *LR* logistic regression, *SVM* support vector machine, *XGBoost* extreme gradient boosting

ML algorithms are a relevant and promising theme not only for neurosurgery but also for the entirety of medicine, seeing that it might provide a stronger approach to guide clinical decision-making. Despite the use of LR in predicting DCI, its use in analyzing complex and large datasets is limited, as this method assumes a linear association between predictors and the outcome variable, which is not always applicable [10]. Although ML algorithms, such as RF and ANN, have been proposed as an alternative to LR for prognosticating mortality in sepsis, their application in DCI predictions needs more research. Most previous studies investigating the use of ML for predicting DCI have not conducted a comprehensive analysis encompassing comparisons with LR. As an illustration, De Jong et al. [6] and Tanioka et al. [11] analyzed the diagnostic of DCI after SAH using ML algorithms but they did not provide a comparison with the performance of LR. To the best of our knowledge, this is the first systematic review and meta-analysis directly comparing LR and ML forecasting performance in the context of DCI following SAH.

The discussion of specific scenarios or contexts in which certain models may exhibit greater effectiveness than others is critical. The choice of ML or regression models should align with the characteristics and variables available in the study setting. For instance, one study identified CT value of SAH, white blood cell count, neutrophil count, CT value of cerebral edema, and monocyte count as significant predictors [14]. Similarly, three studies underscored the relevance of specific clinical variables such as clinical scores, aneurysm treatment, middle cerebral artery peak systolic velocities, and the presence of epileptiform abnormalities. In both cases, employing nonlinear ML models, specifically ANN, is more appropriate for accommodating the intricacies associated with these clinical predictors [5, 12, 23]. Conversely, one study highlighted the significance of demographic and laboratory test variables, including age, hemoglobin, sodium, white blood cell count, platelets, and creatinine [7]. Within this particular context, linear regression models offer enhanced suitability in capturing the linear relationships between these variables and the occurrence of DCI. Furthermore, some models employed specific clinical variables such as clinical scores, aneurysm treatment, middle cerebral artery peak systolic velocities, and the presence of epileptiform abnormalities [5, 12, 23]. In such cases, the incorporation of these variables into ML models, particularly ANN, is more appropriate for accommodating the intricacies associated with these clinical predictors.

Despite our significant findings, this study has some limitations related to the heterogeneity of the results, whose potential sources were identified by a meta-regression analysis. First, the selected studies have different designs; some are retrospective [5, 7, 12] whereas others are prospective [13–15]. Second, they have a lot of differences in their patient populations, including variations in the number of patients involved, the type of SAH (aneurysmal or nonaneurysmal), and the location of medical centers involved, which can impact heterogeneity through differences in inclusion criteria, characteristics of populations, health care systems, environmental factors, and treatment protocols. Third, there is a significant difference in the time point of DCI prediction, DCI diagnostic criteria, and the covariates used for it among the studies.

Notwithstanding these limitations, the combined analysis of nearly 2000 patients from six studies represents a considerable increase in statistical power when compared with individual studies, as it provides a more comprehensive and robust evaluation of the predictive performance of LR and ML algorithms for predicting DCI after SAH. To advance the application of predictive models for DCI, various research directions should be explored. Firstly, it is imperative to standardize the time frame for prediction and the selection of variables across studies, ensuring comparability and reproducibility of findings. Additionally, the timing of DCI prediction should be reported, as this information can provide valuable insights into the temporal sensitivity of the included features.

Table 6 Comparison of ml models strengths and weaknesses

Model	Strengths	Weaknesses
ANN	Ability to capture complex patterns and nonlinear relationships effectively Adaptability to various types of data and problem domains Good performance with large datasets	Computationally expensive, especially with complex architectures and large datasets Requires careful tuning of hyperparameters Interpretability can be challenging due to the black-box nature of the model
DT	Easy to understand and interpret Can handle nonlinear relationships and interactions Robust against outliers	Prone to overfitting, especially with complex trees Not suitable for high-dimensional datasets Lack of robustness to small changes in the data
EC	Improved prediction accuracy compared to individual classifiers Robustness to noise and outliers Effective in handling complex relationships and capturing nonlinear patterns Reduction of bias and variance in predictions Versatility and flexibility in combining different classifiers	Increased complexity and computational requirements Interpretability can be challenging due to the combination of multiple models Selection and configuration require careful consideration Potential risk of overfitting if not properly controlled Lack of transparency in understanding how individual classifiers contribute to the ensemble's decision
GBTM	Identifies distinct subgroups within a population Models individual trajectories over time Provides insights into population dynamics	Assumes predefined number of trajectory groups Sensitivity to initial group assignment Limited interpretability of the resulting trajectory groups
KNN	Simple and intuitive algorithm Can capture complex relationships in the data No assumptions about the underlying data distribution	Computationally expensive during prediction, especially with large datasets Sensitivity to the choice of k (number of neighbors) Not efficient in high-dimensional spaces
MCLP	Nonlinear modeling capability Universal approximation property Robustness to noisy data	Overfitting potential, especially with limited data Computationally expensive training and inference Lack of interpretability due to the black-box nature
RF	Ability to handle nonlinear relationships and interactions effectively Robust against outliers and noise Can handle missing values without significant impact on performance	Lack of interpretability compared to simpler models like LR Computationally more expensive, especially with a large number of trees Requires careful tuning of hyperparameters
SVM	Effective in capturing nonlinear relationships through the use of different kernel functions Robust against overfitting Works well with small to medium-sized datasets	Computationally expensive during training, especially with large datasets Requires careful tuning of hyperparameters and selection of kernel functions Interpretability is limited
XGBoost	Excellent performance in capturing nonlinear relationships and interactions Effective handling of missing values and outliers Efficient computation and scalability	Requires careful tuning of hyperparameters Prone to overfitting if not properly regularized Interpretability is relatively low compared to simpler models

ANN artificial neural network, *EC* ensemble classifier, *GBTM* group-based trajectory modeling, *KNN* K-nearest neighbor, *LASSO* least absolute shrinkage and selection operator, *MCLP* multilayer perceptron, *RF or RFC* random forest, support vector machine, *XGBoost* extreme gradient boosting

The enhancement of statistical power and the broadening of the generalizability of developed models can be achieved through the augmentation of sample size in future studies. Moreover, the exploration of additional predictors and variables beyond the currently employed demographic, clinical, and radiographic characteristics has the potential to offer valuable insights into the underlying mechanisms and risk factors of DCI.

Lastly, we highly recommend conducting clinical interventional trials following the Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence and Consolidated Standards of Reporting Trials—Artificial Intelligence guidelines to assess the effectiveness of ML-based decision support systems.

Conclusions

For patients with SAH, ML models performed slightly better than LR models. These findings imply that ML algorithms have the potential to surpass traditional statistical methods in this context. However, further research is required to explore the different techniques of ML and whether patients' characteristics, data, and diagnostic criteria used for detection could affect their performance. Overall, this meta-analysis provides valuable insights into the use of artificial intelligence in medical research and highlights the potential benefits of these methods for improving patient outcomes.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1007/s12028-023-01832-z.

Author details
[1] School of Medicine, University of São Paulo, São Paulo, SP, Brazil. [2] Department of Neurology, University of São Paulo, São Paulo, SP, Brazil. [3] Division of Neurosurgery, University of São Paulo, São Paulo, SP, Brazil.

Author Contributions
LSS: selection of studies, data collection, statistical analysis, and manuscript writing. JBCD: selection of studies, data collection, risk of bias, and manuscript writing. NNR: supervision and manuscript review. MJT: supervision and manuscript review. EGF: supervision and manuscript review. JPMT: statistical analysis, supervision, and manuscript review. The final manuscript was approved by all authors.

Source of Support
This work was not supported by any grant.

Conflict of interest
The authors declare that they have no conflict of interest.

Ethical Approval/Informed Consent
All authors confirm adherence to ethical guidelines—no data, including images, have been falsified or manipulated to support our conclusions—and the use of a reporting checklist for systematic reviews and meta-analysis.

Publisher's Note

Received: 24 March 2023 Accepted: 4 August 2023
Published: 5 September 2023

References

1. Macdonald RL. Delayed neurological deterioration after subarachnoid haemorrhage. Nat Rev Neurol. 2014;10(1):44–58. https://doi.org/10.1038/nrneurol.2013.246.
2. Francoeur CL, Mayer SA. Management of delayed cerebral ischemia after subarachnoid hemorrhage. Crit Care. 2016;20(1):277. https://doi.org/10.1186/s13054-016-1447-6.
3. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996;49(11):1225–31. https://doi.org/10.1016/s0895-4356(96)00002-9.
4. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255–60. https://doi.org/10.1126/science.aaa8415.
5. Ramos LA, van der Steen WE, Sales Barros R, Majoie CBLM, van den Berg R, Verbaan D, Vandertop WP, Zijlstra IJAJ, Zwinderman AH, Strijkers GJ, Olabarriaga SD, Marquering HA. Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage. J Neurointerv Surg. 2019;11(5):497–502. https://doi.org/10.1136/neurintsurg-2018-014258.
6. De Jong G, Aquarius R, Sanaan B, Bartels RHMA, Grotenhuis JA, Henssen DJHA, Boogaarts HD. Prediction models in aneurysmal subarachnoid hemorrhage: forecasting clinical outcome with artificial intelligence. Neurosurgery. 2021;88(5):E427–34. https://doi.org/10.1093/neuros/nyaa581.
7. Savarraj JPJ, Hergenroeder GW, Zhu L, Chang T, Park S, Megjhani M, Vahidy FS, Zhao Z, Kitagawa RS, Choi HA. Machine learning to predict delayed cerebral ischemia and outcomes in subarachnoid hemorrhage. Neurology. 2021;96(4):e553–62. https://doi.org/10.1212/WNL.0000000000011211.
8. Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Alzheimer's disease neuroimaging initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage. 2011;55(3):856–67. https://doi.org/10.1016/j.neuroimage.2011.01.008.
9. Syeda-Mahmood T, et al. Identifying patients at risk for aortic stenosis through learning from multimodal data. Medical Image computing and computer-assisted intervention—MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science(), vol 9902. Springer, Cham. https://doi.org/10.1007/978-3-319-46726-9_28
10. Ranganath R, Gerrish S, Blei DM. Deep survival analysis. In: Proceedings of the 32nd international conference on machine learning, 2016; pp. 2079–2088. https://cims.nyu.edu/~rajeshr/papers/Ranganath_DeepSurvival2016.pdf
11. Tanioka S, Ishida F, Nakano F, Kawakita F, Kanamaru H, Nakatsuka Y, Nishikawa H, Suzuki H, pSEED group. Machine learning analysis of matricellular proteins and clinical variables for early prediction of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage. Mol Neurobiol. 2019;56(10):7128–35. https://doi.org/10.1007/s12035-019-1601-7.
12. Alexopoulos G, Zhang J, Karampelas I, Khan M, Quadri N, Patel M, Patel N, Almajali M, Mattei TA, Kemp J, Coppens J, Mercier P. Applied forecasting for delayed cerebral ischemia prediction post subarachnoid hemorrhage: methodological fallacies. Inform Med Unlock. 2022;28:100817. https://doi.org/10.1016/j.imu.2021.100817.
13. Chen HY, Elmer J, Zafar SF, Ghanta M, Moura Junior V, Rosenthal ES, Gilmore EJ, Hirsch LJ, Zaveri HP, Sheth KN, Petersen NH, Westover MB, Kim JA. Combining transcranial doppler and EEG data to predict delayed cerebral ischemia after subarachnoid hemorrhage. Neurology. 2022;98(5):e459–69. https://doi.org/10.1212/WNL.0000000000013126.
14. Hu P, Li Y, Liu Y, Guo G, Gao X, Su Z, Wang L, Deng G, Yang S, Qi Y, Xu Y, Ye L, Sun Q, Nie X, Sun Y, Li M, Zhang H, Chen Q. Comparison of conventional logistic regression and machine learning methods for predicting delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage: a multicentric observational cohort study. Front Aging Neurosci. 2022;17(14):857521. https://doi.org/10.3389/fnagi.2022.857521.
15. Megjhani M, Terilli K, Weiss M, Savarraj J, Chen LH, Alkhachroum A, Roh DJ, Agarwal S, Connolly ES Jr, Velazquez A, Boehme A, Claassen J, Choi HA, Schubert GA, Park S. Dynamic detection of delayed cerebral ischemia: a study in 3 centers. Stroke. 2021;52(4):1370–9. https://doi.org/10.1161/STROKEAHA.120.032546.
16. Schmidt J, Marques MRG, Botti S, et al. Recent advances and applications of machine learning in solid-state materials science. N Engl J Med. 2019;5(1):83. https://doi.org/10.1038/s41524-019-0221-0.
17. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9. https://doi.org/10.1056/NEJMp1606181.PMID:27682033;PMCID:PMC5070532.
18. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol. 2010;6:109–38. https://doi.org/10.1146/annurev.clinpsy.121208.131413.
19. Saadatfar H, Khosravi S, Joloudari JH, Mosavi A, Shamshirband S. A new K-nearest neighbors classifier for big data based on efficient data pruning. N Engl J Med. 2020;8(2):286. https://doi.org/10.3390/math8020286.
20. Reifman J, Feldman EE. Multilayer perceptron for nonlinear programming. Comput Oper Res. 2002;29(9):1237–50. https://doi.org/10.1016/S0305-0548(01)00027-2.
21. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 2007;25(8):25. https://doi.org/10.1186/1471-2105-8-25.
22. Cortes C, Vapnik V. Support vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/BF00994018.
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785