Original Research Article

EUROPEAN STROKE JOURNAL

European Stroke Journal
2024, Vol. 9(4) 1053–1062

DOI: 10.1177/23969873241253366
journals.sagepub.com/home/eso
Sage

Artificial intelligence to predict individualized outcome of acute ischemic stroke patients: The SIBILLA project

Pietro Caliandro[1], Jacopo Lenkowicz[2], Giuseppe Reale[3], Simone Scaringi[4], Aurelia Zauli[1], Christian Uccheddu[4], Simone Fabiole-Nicoletto[4], Stefano Patarnello[2], Andrea Damiani[2], Luca Tagliaferri[5], Iacopo Valente[6], Marco Moci[7], Mauro Monforte[1], Vincenzo Valentini[8] and Paolo Calabresi[1,7]

Abstract

Introduction: Formulating reliable prognosis for ischemic stroke patients remains a challenging task. We aimed to develop an artificial intelligence model able to formulate in the first 24 h after stroke an individualized prognosis in terms of NIHSS.

Patients and methods: Seven hundred ninety four acute ischemic stroke patients were divided into a training (597) and testing (197) cohort. Clinical and instrumental data were collected in the first 24 h. We evaluated the performance of four machine-learning models (Random Forest, *K*-Nearest Neighbors, Support Vector Machine, XGBoost) in predicting NIHSS at discharge both in terms of variation between discharge and admission (regressor approach) and in terms of severity class namely NIHSS 0–5, 6–10, 11–20, >20 (classifier approach). We used Shapley Additive exPlanations values to weight features impact on predictions.

Results: XGBoost emerged as the best performing model. The classifier and regressor approaches perform similarly in terms of accuracy (80% vs 75%) and f1-score (79% vs 77%) respectively. However, the regressor has higher precision (85% vs 68%) in predicting prognosis of very severe stroke patients (NIHSS > 20). NIHSS at admission and 24 hours, GCS at 24 hours, heart rate, acute ischemic lesion on CT-scan and TICI score were the most impacting features on the prediction.

Discussion: Our approach, which employs an artificial intelligence based-tool, inherently able to continuously learn and improve its performance, could improve care pathway and support stroke physicians in the communication with patients and caregivers.

Conclusion: XGBoost reliably predicts individualized outcome in terms of NIHSS at discharge in the first 24 hours after stroke.

Keywords

Prognosis, acute ischemic stroke, artificial intelligence (AI), machine learning (ML), outcome prediction

Date received: 12 January 2024; accepted: 21 April 2024

[1]Unit of Neurology, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
[2]Real World Data Facility, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
[3]Unit of High Intensity Neurorehabilitation, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
[4]Ammagamma s.r.l. Via Sant'Orsola 33, Modena, Italy
[5]Unit of Radiotherapy, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
[6]Unit of Interventional Neuroradiology, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
[7]Department of Neurosciences, Università Cattolica del Sacro Cuore, Rome, Italy
[8]Department of Oncology and Radiology, Ospedale Isola Tiberina-Gemelli Isola, Rome, Italy

Corresponding author:
Pietro Caliandro, Unit of Neurology, Fondazione Policlinico Universitario A. Gemelli IRCCS, L.go F. Vito, 1-00168, Rome 00135, Italy.
Email: pietro.caliandro@policlinicogemelli.it

Introduction

Ischemic stroke is a leading cause of disability and mortality worldwide.[1] Moreover, the socio-economic burden associated with stroke is substantial, encompassing medical expenses, rehabilitation costs, lost productivity, and caregiver responsibilities. In Europe, projections show that with a "business as usual" approach, the burden of stroke will not decrease in the next decade. An important contributing factor is that the number of older persons in Europe is rising, with a projected increase of 35% between 2017 and 2050. The total cost of stroke in the EU of an estimated 45–60 billion euros in 2015 is set to rise, including both healthcare and non-healthcare costs.[2] Ischemic stroke is a time-dependent pathology that requires a fast and effective management of the acute phase, because revascularization therapies, accessible in the first hours, strongly reduce the risk of death and the severity of post-stroke disability. The current guidelines rely on data that come from randomized controlled trials with specific inclusion and exclusion criteria.[3–9] However, from daily clinical practice, we know that real-world patients might not perfectly fit such criteria. Therefore, in individual patients treatments and outcomes might differ from what described in the current literature. In this view, prognostic models based on real-world data might contribute to identify factors influencing outcomes in each individual stroke patient and help stroke physicians to tailor assistance in the different stages of stroke pathway. Nevertheless, the current prognostic scores, such as ASTRAL,[10] DRAGON, iScore, PLAN, and CoRisk often require information difficult to collect in the very early hours after stroke or only available at discharge (e.g. etiology, medical complications, etc) and individualized, automated and fast prognostic models are still lacking. Artificial Intelligence (AI) based algorithms can evaluate otherwise overwhelming data and can provide useful individualized outcome prediction.[11–13] Moreover, AI can provide reliable evidence based on the continuous and auto-correcting analysis of big real-world data.[14–17] In fact, Machine Learning makes minimal assumptions on systems generating data; it can be effective even when data are collected without a controlled study design and in case of complicated non-linear interactions.[18] With the present study, our aims were: (1) to develop and train an AI based model able to formulate in the first 24 h after stroke an automated and individualized stroke prognosis in terms of NIHSS score at discharge from the hospital; (2) to test the efficacy of such model.

Materials and methods

Patient population

This is an observational study conducted on acute ischemic stroke patients admitted to Fondazione Policlinico Gemelli between July 2019 and June 2023. The cohort comprehended all consecutive patients who were discharged from the hospital with a diagnosis of acute ischemic stroke, identified through the ICD-9 code (International Statistical Classification of Diseases). Patients with haemorrhagic stroke and patients with stroke mimics were excluded from the study. There were no other inclusion or exclusion criteria, meaning any limit of age, etiology, clinical severity or type of acute revascularization treatment. This approach guarantees to assess a real-world scenario including a large range of patients' characteristics without selection according to anamnestic features, clinical severity, findings of laboratory exams, neuroimaging findings and type of acute treatment (thrombolysis/endovascular treatment). Moreover, missing data in any of the collected variables was not an exclusion criteria as our approach preserves the real-world setting where AI algorithms can add more trustworthy insight even with incomplete datasets.[19] The study was approved by the ethics committee of Fondazione Policlinico Universitario A. Gemelli-IRCCS (prot n. 0020981/21).

Clinical and instrumental data collected

Clinical and instrumental data were collected in the first 24 h after stroke. Data were available in databases of the hospital and stored as text in clinical and radiological reports or as structured data. At arrival at the emergency department we collected personal information (age and gender) and medical history of the patients (history of hypertension, diabetes mellitus, atrial fibrillation, cardiac diseases, anemia, dementia, previous transient ischemic attack (TIA)/ischemic/haemorrhagic stroke, ongoing antiplatelet/anticoagulant therapy before the index event as dichotomic variables) from the text of clinical reports, as well as clinical characteristics of the index ischemic event such as known onset of symptoms (dichotomic variable) and clinical severity measured with NIHSS, along with vital parameters (systolic and diastolic blood pressure, heart and breathing rate). Moreover, at the emergency department we collected relevant information as dichotomic variables from the first brain CT and brain CT-angiography such as presence of acute ischemic lesion and/or chronic leukoencephalopathy occlusion of intra/extracranial large vessels, site of occlusion, evidence of good/poor collateral blood flow qualitatively defined by the radiologist in the radiological report, execution of perfusion studies. Finally, we collected the reperfusion treatment type (thrombolysis/endovascular treatment), if any, and the reperfusion grade by TICI score for endovascular treated patients.

We also evaluated clinical severity within 24 h after the admission (NIHSS or Glasgow Coma Scale in case of disorder of consciousness). Table S1 of Supplemental Materials shows all variables acquired and depict how each variable is encoded. Specifically, dichotomic variables were classified as present and encoded as 1 when the text

mining procedure identified it in the medical/instrumental report. When the search produced no matching finding, the variable was set to 0. Therefore, for each variable, "0" comprises both the absence of that feature in a specific patient or that the text mining procedure was not able to detect it.

Machine learning models construction

Data preprocessing. The dataset included acute ischemic stroke patients admitted to Policlinico Gemelli between the period from July 2019 to June 2023, considering that the management of stroke care pathway was different before July 2019 and remained unchanged during this time-frame. Research studies that aim to provide AI-based predictive models, such as this one, imply a heuristic process, where key aspects of the clinical research, such as the size effect, the desired statistical power and the minimum sample size are not available at the time of the study design. Therefore, an evaluation of the performance of these models on previously unseen data is of the utmost importance. Thus, the cohort was randomly divided into two samples: 75% of the patients were allocated to the machine learning model training set, while the remaining 25% was utilized as testing set for evaluating the performance of the trained models. The primary sources from which features were extracted consist of two main categories: clinical/radiological reports (in textual form), as well as pre-existing tabulated data. To ensure the integrity and uniformity of the dataset, the textual content of clinical documents underwent preprocessing procedures as conversion to lowercase to ensure consistency in text case, removal of punctuation marks to streamline the text, elimination of diacritics and accents to simplify the text representation, rectification of known errors, such as "NHISS" corrected to "NIHSS," among others. We developed a specific ontology for identifying unstructured information from the hospital registries and coding them in structured variables to be used in the AI models. This ontology has been used for text mining and automatically collecting significant anamnestic features, clinical severity and neuroimaging findings from clinical and radiological reports. Text mining techniques such as pattern matching with regular expression and lemmas co-occurrence analysis were employed to make the ontology actionable at the software level. Categorical variables underwent the procedure of one-hot encoding, a technique employed to transform categorical data into a numerical format suitable for machine learning algorithms. Notably, some variables, as TICI score, possess an inherent ordinal nature and underwent transformation into ordinal variables. This adjustment preserved the intrinsic ordering of the variable, thus allowing for a more meaningful representation within the model.

To ensure the accuracy, consistency, and completeness of data used for analysis, a two-part quality control process is implemented. First, daily batch procedures running during off-peak hours check data extracted from hospital operational systems. Each data stream has a designated process owner who tracks and corrects errors. Data quality checks then delve deeper. These checks include ensuring clear definitions (meaning, units, and standardized classifications) for each indicator, identifying missing or duplicate values in key fields (like dates), verifying data consistency (e.g. discharge after admission), and comparing trends with data providers. Finally, the Service Desk rectifies any flagged inconsistencies in the original registration application. The output of this validation process is compared against benchmark distributions taken from historical data and is considered fair if the level of agreement is above a predefined threshold, which is set at the 95% confidence level. Any residual discrepancy or outlier is further investigated on a case-by-case basis with the clinical team in charge of the study. This combination of automated checks and human oversight safeguards the quality of the data used for the analysis.

Definition of the prediction target. The designated target variable in this study is the NIHSS score at the time of discharge, as an indicator of the patient's prognosis and response to therapeutic interventions. Discharge occurred during the acute phase when the diagnostic workflow had been completed and patients had not any complication requiring specific medical and/or surgical interventions. Patients were transferred to a rehabilitation center or at home according to the clinical severity. Patients coming from the spoke hospitals with indication to endovascular treatment were transferred back to the spoke center if no further interventions were required after endovascular procedure. Two different prediction tasks have been developed: a multiclass classifier task and a regressor task. For the multiclass classifier approach NIHSS score at discharge is categorized into four classes: NIHSS 0–5 (absence of symptoms or minor stroke), NIHSS 6–10 (mild stroke), NIHSS 11–20 (severe stroke), NIHSS $>$ 20 (very severe stroke). Regarding the regressor approach, the prediction output is defined as variation of NIHSS score between the value at discharge and at admission and it incorporates the sign (positive for deterioration, negative for improvement). Deceased patients were classified as having a NIHSS score of 43 (60 deceased patients in the training cohort and 19 deceased patients in the testing cohort).

Models selection and evaluation metrics. We evaluated four machine-learning models for both regression and classification tasks: Random Forest (RF), *K*-Nearest Neighbors (kNN), Support Vector Machine (SVM), and XGBoost (XGB). To select the best model we used a fivefold cross validation. For the multiclass classifier approach we adopted f1-score weighted on classes as evaluation metric while for the regression approach we used the Root Mean Squared Error (RMSE). The four machine-learning models underwent distinct training procedures for the classifier and

regressor approaches. The training of the classifier was executed utilizing the softprob (softmax) loss function. Conversely, the training for the regressor was meticulously designed to optimize the squared error.

The best performing models in the training cohort, both in the classifier and regressor approaches, were applied in the testing cohort. We assessed the performance of the models on the testing set using a variety of metrics. For the classification task we considered accuracy, precision, recall, f1-score, macro average (average of the precision, recall, and f1-score across all classes, without considering class imbalance) and weighted average (average of the precision, recall, and f1-score) where each class is weighted by its support (the number of instances in each class). In the regression task, we computed the absolute error distribution over the testing set (5th, 10th, 25th, 75th, 90th Percentiles). The absolute error was also computed by clinically relevant classes of NIHSS at admission, because the impact of the error for these classes is different (e.g. if the model predicts the NIHSS at discharge with an error of 4 points that error is more meaningful for patients with a NIHSS at admission 0–5 than for patients with an admission NIHSS >20). In order to compare the performances of the two tasks (multiclass classifier and regressor), we used the NIHSS variation predicted by the regressor to predict NIHSS numerical value at discharge with the following formula:

$$\widehat{NIHSS}_{discharge} = NIHSS_{admission} + \widehat{\Delta}_{pred}$$

Then we assigned the obtained NIHSS value to the same NIHSS classes used for the classification task. In this way we converted the findings from the regressor into classes of severity.

We used Shapley Additive exPlanations (SHAP) values to get a consistent and objective explanation of how each feature impacts the models' prediction.

The code used to perform the study has been developed using Python 3.8. All the libraries used for the Artificial Intelligence modules are open source, available online and listed in the Appendix 1 of Supplemental Materials.

Statistical analysis

We assessed normality of continuous variables distribution by the Shapiro-Wilk test. Differences between groups regarding clinical characteristics and duration of hospital stay were calculated using Mann-Whitney *U* test. For each variable, Agresti-Caffo test was conducted to compare distribution of the variables between the training and the testing cohort. The Pearson's correlation coefficient was used as a measure of linear correlation between two variables. Statistical significance was defined as $p < 0.05$.

Results

A total of 794 acute ischemic stroke patients were included in the study. The training and testing cohorts (597 and 197 patients respectively) did not differ in terms of duration of hospital stay (respectively 7 ± 8 days and 8 ± 11 days), distribution of the variables (Supplemental Materials Table S1), in the clinical severity as measured by NIHSS [Figure 1] and number of deaths (60 and 19 respectively). Figure 2 shows the correlation matrix between features in the training cohortF2. In the training cohort XGB resulted as the best performing model for both the classifier (weighted f1-score: 75.3%) and regressor approaches (RMSE: 8.91), so it was applied in the testing cohort to assess its performance on previously unseen data. Table 1 shows the performance metrics for the classifier and regressor in the training cohort.

Classifier performance in the testing cohort

In the testing cohort the classifier approach showed a weighted f1-score of 79% and an accuracy of 80% (Figure 3(a) shows the confusion matrix with evaluation metrics) F3. For the classifier the main difficulty was to predict the 6–10 NIHSS class (f1-score 46%). Figure S1 of Supplemental Materials shows the confusion matrix with evaluation metrics excluding dead patients from the analysis. The most important features, computed using SHAP values, are the following: NIHSS at admission, 24 h NIHSS, 24h GCS, TICI score, heart rate, presence of acute ischemic lesion at the first brain CT scan (Figure 4(a) shows the SHAP plot)F4. For the global feature importance interpretation, it is possible to state that a high NIHSS value (both at admission and within 24 h) tends to increase the probability of a higher NIHSS at the discharge, as well as a low GCS value within 24 h increases the probability of having a high NIHSS at discharge. Moreover, higher heart rate values at admission as well as the presence of an acute ischemic lesion at the first brain CT-scan increases the probability of having higher NIHSS at discharge, while higher values of TICI score after the endovascular procedures decreases the probability of a high NIHSS at discharge. However it is important to state that the single prediction on the single patient depends on the combination of values of all variables.

Regressor performance in the testing cohort

In the testing cohort, the regressor approach showed a median absolute error of 2 in predicting the variation between NIHSS at discharge and at admission (Figure 5 shows the distribution of the absolute error)F5. Table S2 of Supplemental Materials shows the distribution of absolute error by NIHSS class at admission. Figure 3(b) shows the confusion matrix for the regressor results. In the figure, the

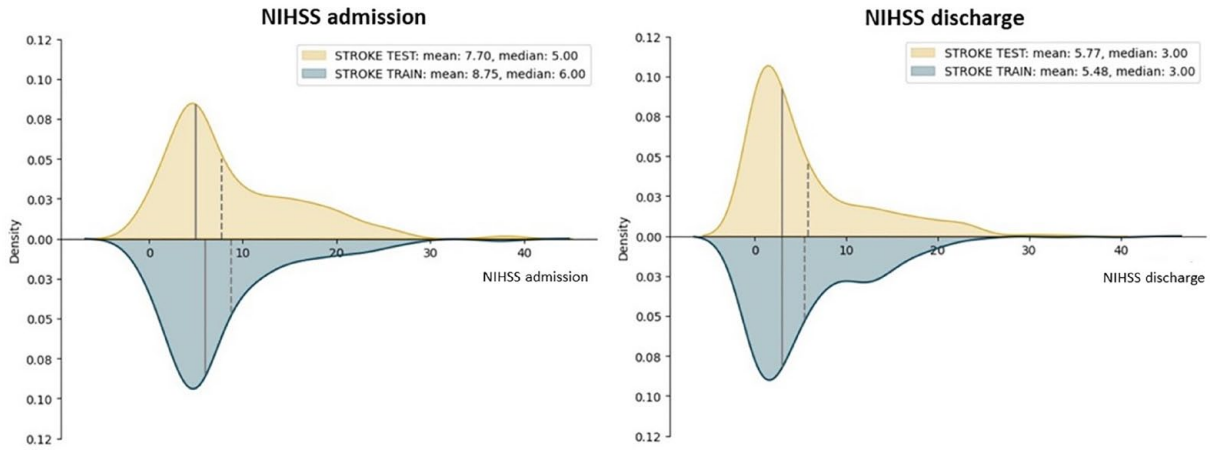


Figure 1. NIHSS distribution at admission and at discharge in training and testing cohorts.

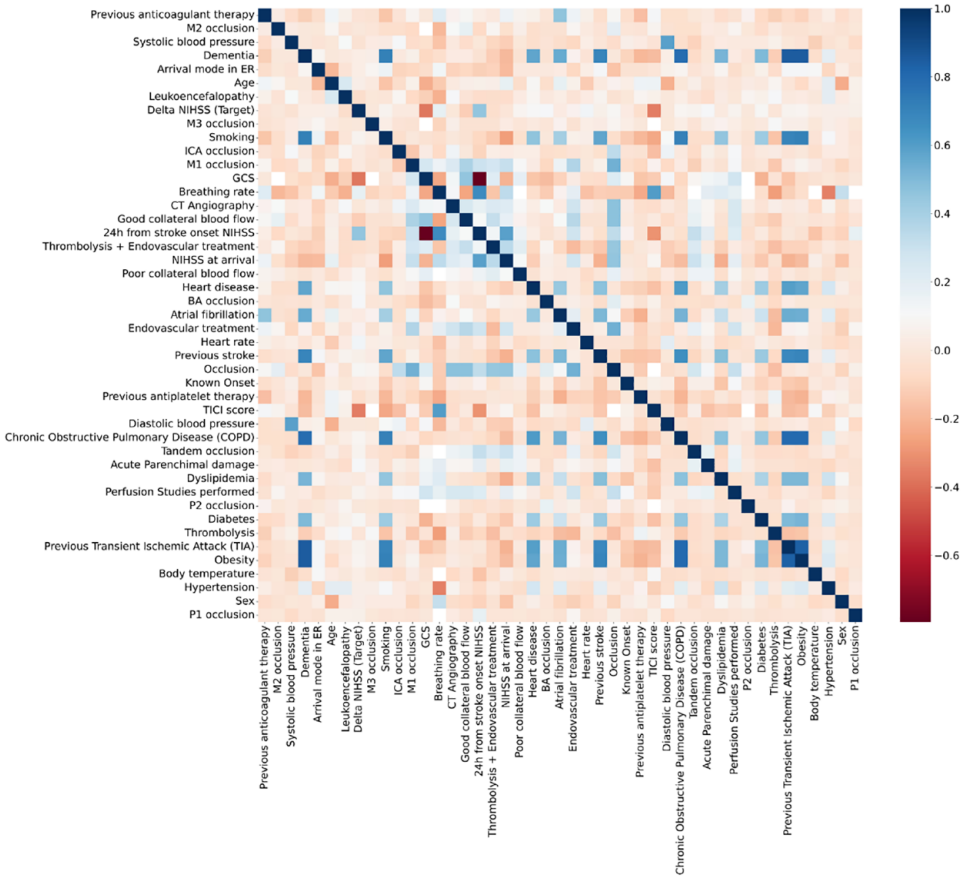


Figure 2. Correlation matrix between features in the training cohort.

predicted NIHSS class at discharge has been calculated by using the predicted NIHSS variation. In the testing cohort the regressor approach showed a weighted f1-score of 77% and an accuracy of 75%. For the regressor approach the main difficulty was to identify the 6–10 NIHSS class (f1-score 43%). Figure S1 of Supplemental Materials shows the confusion matrix with evaluation metrics excluding dead patients from the analysis.

The global feature importance plot (Figure 4(b)), computed using SHAP values, shows that a high NIHSS value

Table 1. Performance metrics for each machine learning model for classifier and regressor in the cross-validation evaluation. f1-score weighted on classes was used for the classifier and the Root Mean Squared Error (RMSE) for the regression approach.

Model	Classifier	Regressor
Random Forest	0.702	10.1
K-NN	0.687	12.90
SVM	0.623	14.32
XGBoost	0.753	8.91

at admission generally is linked to a prediction of improvement (negative delta). Moreover, a low GCS and high NIHSS value at 24 h are linked to a prediction of clinical worsening (positive NIHSS delta), as well as a higher heart rate and the presence of acute ischemic lesion at the first brain CT-scan. On the other hand, high TICI values are associated to a prediction of improvement (negative delta).

Comparison between classifier and regressor performance

The two approaches perform similarly in terms of both accuracy (80% for the classifier and 75% for the regressor) and f1-score (79% for the classifier and 77% for the regressor). The f1 score is essentially the same for the two approaches even when we compare the f1 score values within each class (Figure 3).

Looking more deeply, we see that the two approaches performed better for NIHSS class 0–5 and 11–20, (Figure 3), while they performed worse for the NIHSS class 6–10, that was the most difficult to predict with similar f1 scores (46% for the classifier and 43% for the regressor). However, for the NIHSS class 6–10 the regressor task performs better than the classifier in terms of recall (60% vs 45%), that is, the regressor correctly identifies 60% of the patients with a true value of NIHSS of 6–10. The majority of patients in the remaining 40% is predicted to have a NIHSS value in the class of clinical severity 0–5 (Figure 3). The regressor performs worse than the classifier in terms of precision (33% vs 47%). This means that among patients predicted to have a NIHSS 6–10, 33% actually belongs to that class. However, the majority of the remaining 67% has a NIHSS value of 0–5. In clinical terms, this means that these patients will have a better outcome than predicted.

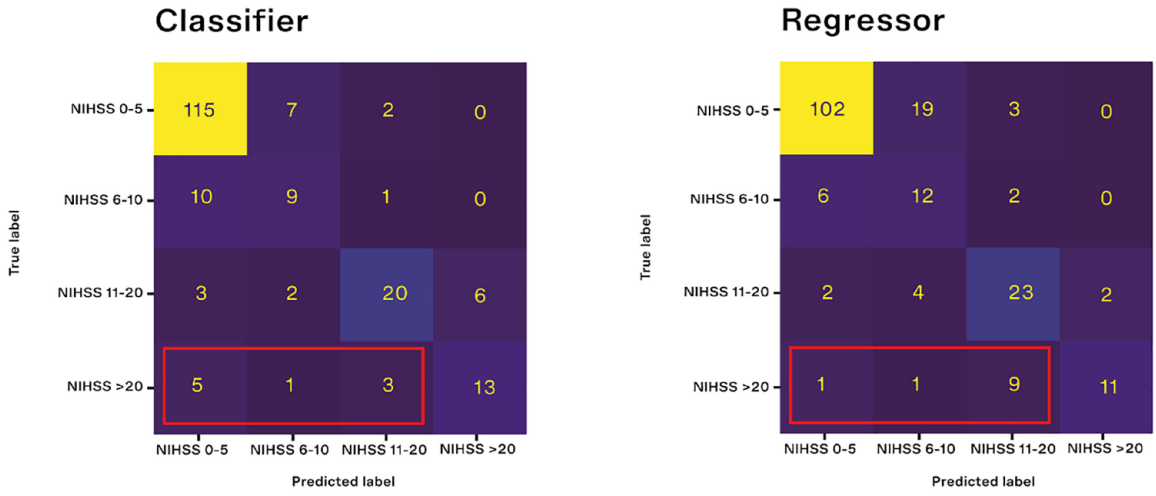
If we look at the NIHSS class >20, the two approaches perform equally in terms of f1 score (63%). The regressor has higher precision (85% vs 68%), meaning that among patients predicted to have a NIHSS > 20, 85% actually belongs to that class. The remaining 15% of patients whose outcome is not correctly predicted by the regressor actually belong to the nearest lower class of severity (NIHSS 11–20). On the other hand, the regressor performs slightly worse than the classifier in terms of recall (50% vs 59%), meaning that among all patients with a NIHSS > 20, the regressor correctly identifies 50% of them. However, as reported in Figure 3(b), among patients that actually have a NIHSS > 20 the majority of the incorrectly predicted patients are predicted as having a NIHSS value of 11–20 by the regressor, while they are predicted to have NIHSS between 0 and 5 by the classifier.

Moreover, since important clinical changes could occur in the first 24 h after the index event, we evaluated the performance of XGBoost model (both classifier and regressor approaches) using only NIHSS at 24 h as predictor of NIHSS at discharge and verified that this last approach is less performing than a full model using a more comprehensive approach with all the variables we selected (Table S3 of Supplemental Materials).

Discussion

This study demonstrated that XGBoost algorithm was able to reliably predict patients' NIHSS at discharge based on real-world data collected within the first 24 h after ischemic stroke onset. The classifier and regressor approaches perform similarly in terms of accuracy (80% for the classifier and 75% for the regressor) and f1-score (79% for the classifier and 77% for the regressor). However, we have to consider that when the regressor fails to correctly predict the NIHSS of a patient with an actual NIHSS > 20, it generally predicts a NIHSS value of 11–20, so it still predicts a severe clinical condition. On the other hand, when the classifier fails to correctly predict the NIHSS of a patient with an actual NIHSS > 20 at discharge, it more probably predicts a NIHSS 0–5, that is a minor stroke, being this more optimistic prediction obviously unacceptable from a clinical point of view. In light of this, we should prefer the regressor approach over the classifier, even if their global performance is similar. In fact, while the number of patients not correctly classified by the regressor in the class NIHSS > 20 is higher, the prediction error is less serious from a clinical perspective. In other words, we could say that the regressor prediction is more pessimistic than that of the classifier approach in those cases of misclassified patients with actual NIHSS > 20.

With respect to the interpretation of global importance of the variables in making prediction, in both the approaches SHAP plots demonstrate that a high NIHSS value at 24 h, a low GCS score at 24 h and higher heart rate values at admission, as well as the presence of an acute ischemic lesion at the first brain CT increase the probability of having higher NIHSS at discharge, while higher values of TICI score after the endovascular procedures decrease the probability of a higher NIHSS at discharge. The only difference between these two approaches was the NIHSS value at admission. In the classifier approach, high NIHSS values at admission



	precision	recall	f1-score	support
NIHSS 0-5	0.86	0.93	0.89	124
NIHSS 6-10	0.47	0.45	0.46	20
NIHSS 11-20	0.77	0.65	0.70	31
NIHSS >20	0.68	0.59	0.63	22
accuracy			0.80	197
macro avg	0.70	0.65	0.67	197
weighted avg	0.79	0.80	0.79	197

	precision	recall	f1-score	support
NIHSS 0-5	0.92	0.82	0.87	124
NIHSS 6-10	0.33	0.60	0.43	20
NIHSS 11-20	0.62	0.74	0.68	31
NIHSS >20	0.85	0.50	0.63	22
accuracy			0.75	197
macro avg	0.68	0.67	0.65	197
weighted avg	0.80	0.75	0.77	197

Figure 3. Evaluation metrics and confusion matrices between predicted and true NIHSS for classifier (a) and regressor (b). For the regressor approach, we have converted the predicted NIHSS variation into predicted NIHSS classes to make comparable the findings from classifier and regressor approaches.

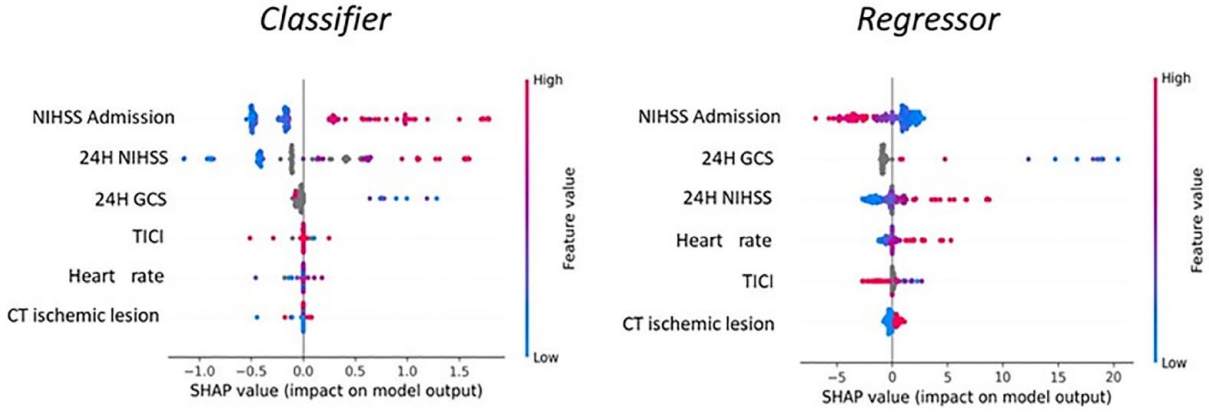


Figure 4. SHAP plots describing the main features ordered according to their importance in making For Peer Review prediction as classifier (a) and regressor (b).

shift the prediction toward high NIHSS classes at discharge, while low NIHSS values at admission shift the prediction toward low NIHSS classes at discharge. On the other hand, the regressor task works in a less intuitive way. In fact, low NIHSS values at admission generally shift the prediction toward a positive variation of NIHSS between discharge and admission meaning neurological deterioration. This is due to the fact that, in case of clinical worsening of a patient with low NIHSS at admission, the amount of positive variation can theoretically be very wide for example, from NIHSS 5 at admission to 43 at discharge (the highest possible variation is +38). On the other hand, in case of clinical improvement of patients with low NIHSS at admission (negative NIHSS variation), necessarily the amount of variation must be little, for example, from NIHSS 5 at admission a patient cannot improve more than 5 points, meaning

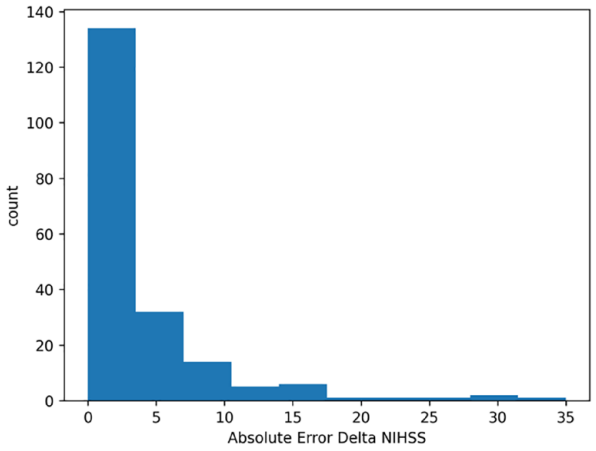


Figure 5. Absolute error distribution of the predicted variation between NIHSS at discharge and at admission.

that the highest possible variation is −5. Therefore, low NIHSS at admission cannot be associated to great negative variation but only to great positive variations. The same interpretation works if we consider high NIHSS values at admission: high NIHSS values at admission tend to shift the prediction toward a negative variation. However, it is crucial to bear in mind that the outcome prediction on the single patient depends on the combination of all the variables collected and AI-modeled from real-world big data, making the prediction tailored and unique for each and every patient.

To our knowledge, while there are other research studies that used AI-based models to predict long term clinical outcomes,[11–13,20] this is the first study aimed to predict a short-term neurological outcome in terms of NIHSS. In fact, almost all of the studies published up to this point predicted the mRS at 90 days.[11–13,20] Lin et al.[12] excluded patients who died before discharge, Heo et al.[11] excluded patients who received recanalization treatment, while Xie et al.[13] excluded patients with vertebrobasilar occlusions. Moreover, in building their models, these Authors considered some information which could be available only after a diagnostic workflow performed during the overall hospital stay, as clinical and instrumental information aimed to assess stroke etiology according to the TOAST classification[11,20] or clinical severity at discharge in terms of NIHSS, Barthel index, and mRS scores.[12] Our approach instead preserves the real-world setting avoiding selection of patients according to their clinical severity, treatment options and anterior/posterior circulation involvement and allow a prognostic prediction based only on features acquired in the first 24 h, the most crucial. Other studies focused on predicting long-term clinical outcome, especially in order to improve selection of patients for endovascular treatment.[21]

In our opinion, the possibility to predict a short-term outcome such as NIHSS at discharge adds a new value to the potential application of this technology in a clinical context. In fact, it could be really useful to have a tool that could accurately predict the clinical pathway of the patients. First of all, it could support stroke physicians in communication with patients and caregivers in the first 24 h. In fact, it is known that stroke physicians' personal expertise and judgment are not sufficiently accurate in making reliable prognosis.[22,23] In our opinion, a tool which bases its predictions on a large amount of data, constantly evolving and performing better, could be a valuable aid in answering prognostic questions. Moreover, we have to consider that NIHSS at discharge is strongly related to mRS at 3 months.[24] Indeed, NIHSS in the acute phase, specifically at 7 days, has been proposed as surrogate outcome of mRS at 90 days.[25] Therefore our approach, which in the first 24 h after stroke makes a prediction of NIHSS at discharge, gives important insights on the level of disability at 3 months and contributes to depict a tailored clinical evolution for each and every patient. Moreover, since the study was conducted in a hub hospital of a stroke network, patients arrive at our center both directly as first aid and from spoke hospitals in order to potentially undergo revascularization procedures. In light of this, the studied cohorts comprised a wide scenario of stroke patients which is indeed difficult to find in smaller hospitals where revascularization treatments might not be available.

Of course this study presents some limitations. First of all in our approach we aimed to predict clinical severity in terms of NIHSS which however does not score dead patients. This aspect is a well known limit of NIHSS[25] and a defined approach to overcome this issue is not available. Given that dead patients were arbitrarily scored as having a NIHSS value of 43, our models include dead patients in the severity class NIHSS > 20. Conceptually it is challenging to consider dead patients in the same category of very severe patients, but this approach allows to avoid a more serious methodological issue as the selection bias of excluding dead patients from analysis. Secondly, the study was conducted in a single clinical center and therefore the model needs to be externally validated in different clinical sites also at an international level in order to develop a model as generalizable as possible. In this view, the process of integrating data from different hospitals must necessarily consider specificities of the different settings such as different health care system organizations, availability of variables and languages in which the variables are expressed. Following this approach, different machine learning models should be evaluated in order to choose the best performing models in these different contexts. Then, algorithms should undergo a process of re-training and updating to confirm that XGBoost is the best performing model in a wider scenario. Moreover, this process will be useful for the overall improvement of the accuracy of the models. However, there could be potential privacy issues regarding the sharing of sensitive personal data

among different institutions, that could be overcome with a federated learning approach. In this view, the described methodology and the findings obtained with this study could represent a first step toward that ambitious target.

Conclusions

In conclusion, our data demonstrated that in the first 24 h after stroke XGBoost may reliably predict clinical evolution in terms of NIHSS at discharge.

Acknowledgements

None.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research received partial funding from the Italian Ministry for University and Research (MUR) under the Program PON Research and Innovation supporting the development of the artificial intelligence platform Gemelli Generator at Fondazione Policlinico Universitario A. Gemelli IRCCS.

Ethical approval

The Ethics Committee of Fondazione Policlinico Universitario A. Gemelli approved this study (Prot N. 0020981/21).

Informed consent

Informed consent was not sought because the study follows a specific procedure applied in our Institution on all the studies developing models of artificial intelligence. This procedure is specifically cited in the approval of the Ethics Committee (Prot N. 0020981/21).

Guarantor

PC

Contributorship

PC and JL and GR researched literature and conceived the study. PC, JL, GR, SP, MM, AD, LT, IV, VV, PC were involved in protocol development and gaining ethical approval. All authors were involved in features selection. PC, GR, JL, SS, CU, SFN, MM, MM, and AZ were involved in data analysis. PC, GR, JL, SS, AZ, and MM wrote the first draft of the manuscript. All authors critically reviewed and edited the manuscript and approved the final version of the manuscript.

ORCID iDs

P. Caliandro https://orcid.org/0000-0002-1190-4879
M. Moci https://orcid.org/0000-0002-7893-1660

Supplemental material

Supplemental material for this article is available online.

References

1. Feigin VL and Owolabi MO. Pragmatic solutions to reduce the global burden of stroke: a world stroke organization-lancet neurology commission. *Lancet Neurol* 2023; 22: 1160–1206.
2. Luengo-Fernandez R, Violato M, Candio P, et al. Economic burden of stroke across Europe: a population-based cost analysis. *Eur Stroke J* 2020; 5: 17–25.
3. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 Guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American heart association/American stroke association. *Stroke* 2019; 50: e344–e418.
4. Berge E, Whiteley W, Audebert H, et al. European stroke organisation (ESO) guidelines on intravenous thrombolysis for acute ischaemic stroke. *Eur Stroke J* 2021; 6: I–LXII.
5. Thomalla G, Simonsen CZ, Boutitie F, et al. MRI-guided thrombolysis for stroke with unknown time of onset. *New Engl J Med* 2018; 379: 611–622.
6. Ma H, Parsons MW, Christensen S, et al. A multicentre, randomized, Double-Blinded, placebo-controlled phase III study to investigate extending the time for thrombolysis in emergency neurological deficits (EXTEND). *Int J Stroke* 2012; 7: 74–80.
7. Campbell BCV, Ma H, Ringleb PA, et al. Extending thrombolysis to 4.5-9hours and wake-up stroke using perfusion imaging: a meta-analysis of individual patient data from EXTEND, ECASS4-EXTEND and EPITHET. *Lancet* 2019; 394: 139–147.
8. Albers GW, Marks MP, Kemp S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *New Engl J Med* 2018; 378: 708–718.
9. Nogueira RG, Jadhav AP, Haussen DC, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N. Engl. J. Med* 2018; 378: 11–21.
10. Ntaios G, Faouzi M, Ferrari J, et al. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. *Neurology* 2012; 79: 2293–2322.
11. Heo J, Yoon JG, Park H, et al. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke* 2019; 50: 1263–1265.
12. Lin CH, Hsu KC, Johnson KR, et al. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Comput Methods Programs Biomed* 2020; 190: 105381.
13. Xie Y, Jiang B, Gong E, et al. JOURNAL CLUB: use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *AJR Am J Roentgenol* 2019; 212: 44–51.
14. Friedman J, Hastie T and Tibshirani R. The elements of statistical learning. Vol. 1. In: *Springer series in statistics*. Springer, 2001.
15. Murphy KP. *Machine learning: A probabilistic perspective*. Cambridge: MIT Press, 2012.

16. Deo RC. Machine learning in medicine. *Circulation* 2015; 132: 1920–1930.
17. Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. *Big Data* 2019; 6: 54.
18. Bica I, Alaa AM, Lambert C, et al. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther* 2021; 109: 87–100.
19. Aydin Z and Ozturk Z. Performance analysis of XGBoost classifier with missing data. In: The 1st international conference on computing and machine intelligence (ICMI 2021), February 2021.
20. Jang SK, Chang JY, Lee JS, et al. Reliability and clinical utility of machine learning to predict stroke prognosis: comparison with logistic regression. *Stroke* 2020; 22: 403–406.
21. van Os HJA, Ramos LA, Hilbert A, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Front Neurol* 2018; 9: 784.
22. Herzog L, Kook L, Hamann J, et al. Deep learning versus neurologists: functional outcome prediction in LVO stroke patients undergoing mechanical thrombectomy. *Stroke* 2023; 54: 1761–1769.
23. Saposnik G, Cote R, Mamdani M, et al. JURaSSiC: accuracy of clinician vs risk score prediction of ischemic stroke outcomes. *Neurology* 2013; 81: 448–455.
24. Zhang MY, Mlynash M, Sainani KL, et al. Ordinal prediction model of 90-day modified rankin scale in ischemic stroke. *Front Neurol* 2021; 12: 727171.
25. Chalos V, van der Ende NAM, Lingsma HF, et al. National institutes of health stroke scale: an alternative primary outcome measure for trials of acute treatment for ischemic stroke. *Stroke* 2020; 51: 282–290.