Journal of Stroke and Cerebrovascular Diseases 33 (2024) 107514

Contents lists available at ScienceDirect

Journal of Stroke and Cerebrovascular Diseases

journal homepage: www.elsevier.com/locate/jstroke

Explaining predictors of discharge destination assessed along the patients' acute stroke journey

Artem Lensky [a,b,*], Christian Lueck [c], Hanna Suominen [c,d,f], Brett Jones [e], Robin Vlieger [d], Tina Ahluwalia [e]

[a] School of Engineering and Technology, The University of New South Wales, Canberra ACT 2600, Australia
[b] School of Biomedical Engineering, The University of Sydney, NSW, Australia
[c] School of Medicine and Psychology, The Australian National University, ACT, Australia
[d] School of Computing, The Australian National University, ACT, Australia
[e] Department of Neurology, Canberra Hospital, ACT, Australia
[f] Department of Computing, University of Turku, Finland

ARTICLE INFO

Keywords:
Stroke
Discharge destination
Machine learning
Clinical diagnosis
Decision support techniques
Health information systems

ABSTRACT

Introduction: Accurate prediction of outcome destination at an early stage would help manage patients presenting with stroke. This study assessed the predictive ability of three machine learning (ML) algorithms to predict outcomes at four different stages as well as compared the predictive power of stroke scores.

Methods: Patients presenting with acute stroke to the Canberra Hospital between 2015 and 2019 were selected retrospectively. 16 potential predictors and one target variable (discharge destination) were obtained from the notes. *k*-Nearest Neighbour (kNN) and two ensemble-based classification algorithms (Adaptive Boosting and Bootstrap Aggregation) were employed to predict outcomes. Predictive accuracy was assessed at each of the four stages using both overall and per-class accuracy. The contribution of each variable to the prediction outcome was evaluated by the ensemble-based algorithm and using the Relief feature selection algorithm. Various combinations of stroke scores were tested using the aforementioned models.

Results: Of the three ML models, Adaptive Boosting demonstrated the highest accuracy (90%) at Stage 4 in predicting death while the highest overall accuracy (81.7%) was achieved by kNN (*k*=2/City-block distance). Feature importance analysis has shown that the most important features are the 24-hour Scandinavian Stroke Scale (SSS) and 24-hour National Institutes of Health Stroke Scale (NIHSS) scores, dyslipidaemia, hypertension and premorbid mRS score. For the initial and 24-hour scores, there was a higher correlation (0.93) between SSS scores than for NIHSS scores (0.81). Reducing the overall four scores to InitSSS/24hrNIHSS increased accuracy to 95% in predicting death (Adaptive Boosting) and overall accuracy to 85.4% (kNN). Accuracies at Stage 2 (pre-treatment, 11 predictors) were not far behind those at Stage 4.

Conclusion: Our findings suggest that even in the early stages of management, a clinically useful prediction regarding discharge destination can be made. Adaptive Boosting might be the best ML model, especially when it comes to predicting death. The predictors' importance analysis also showed that dyslipidemia and hypertension contributed to the discharge outcome even more than expected. Further, surprisingly using mixed score systems might also lead to higher prediction accuracies.

Introduction

A stroke is the sudden onset of neurological dysfunction caused by an interruption of blood supply or by a hemorrhage. It can affect the brain, spinal cord, or eye, and, in general, stroke represents the third leading cause of death in Australia.[1,2] From the time of stroke onset, multiple management decisions must be made, including determining whether the patient has had a stroke[3], determining which hospital is the most appropriate place to manage the patient, deciding whether hyperacute therapy is indicated and, if so, which therapy. Accurate prognostic information contributes to managing the expectations of patients and carers and could enhance cost-effectiveness, for example by highlighting

* Corresponding author.
E-mail address: a.lenskiy@unsw.edu.au (A. Lensky).

https://doi.org/10.1016/j.jstrokecerebrovasdis.2023.107514
Received 9 September 2023; Received in revised form 15 November 2023; Accepted 26 November 2023
Available online 16 December 2023

the need for rehabilitation input at an early stage.

Many authors have proposed tools that can assist in patient management[4], for example, the Field Assessment Stroke Triage for Emergency Destination (FAST-ED).[5] Algorithms based on the National Institutes of Health Stroke Scale (NIHSS)[4] and the Scandinavian Stroke Scale (SSS)[6], and Rapid Arterial oCclusion Evaluation (RACE)[7] are based on the best evidence that is currently available in the literature but there may be additional factors which might influence the outcome that are not currently being considered.

Risk factors for stroke are well known. Investigating the contribution of these risk factors to eventual patient outcomes at the time of admission could improve the precision of decision making.[8] Until recently, studying the impact of large numbers of predictors on individual patients' outcomes has not been possible. Machine Learning (ML) now offers the potential to do this and can therefore help in determining which factors do indeed make a difference to the outcome and therefore require consideration when making management decisions. In this way, management decisions can be better tailored to individual patients[9]. In order to achieve this, methods of interpretable machine learning prove useful. In a recent survey[10], the authors discussed SHapley Additive exPlanations and Gradient Class Activation Mapping for tabular and image data as applied to healthcare systems and showed that feature importance analysis provided medical practitioners with additional information that allowed them to make more knowledgeable decisions using machine learning.

In regard to acute stroke patients, ML models have been successfully applied to predict patients' outcomes. Rana et al. used penalized regression with Lasso to predict discharge destinations using electronic administrative records.[11] The authors considered patients with haemorrhage, ischemic stroke, and transient ischemic attacks and looked at four different outcomes: home, rehabilitation, nursing home and in-hospital death. The authors reported the Area under the Receiver operating characteristic Curve (AUC) 0.85 and 0.825 for predicting discharge to rehabilitation and all other outcomes vs. death for ICH. Alpaydin et al., on the other hand, concluded that no single algorithm was able to predict clinical outcomes with an acceptably high level of accuracy.[12]

Teale et al. concluded that their Six Simple Variables (SSV) model performed as well as more sophisticated models in predicting stroke outcomes.[13] SSV is a regression model with the following six predictors: age, living alone, independence in activities of daily living before the stroke, the verbal component of the Glasgow Coma Scale, arm power, and ability to walk. Counsell et al. proposed to predict 30-day survival after stroke and survival in a non-disabled state at 6 months.[14]

König et al. proposed two models based on age and NIHSS score assessed within 6 h after stroke to predict survival and functional independence 3 months after acute stroke.[15,16] The first model correctly predicted 62.9% of the patients who were left disabled or had died and 83.2% of the patients who had completely recovered. The second model correctly predicted 57.9% of the patients who had died and 91.5% of surviving patients. Chen et al focused on predicting the possibility of offering intravenous thrombolysis to acute ischemic stroke patients and proposed the Intravenous Thrombolysis Score that showed a good predictive accuracy.[17]

In a recent study[18], Yang et al., pointed out that, even though NIHSS scores play a key role in assessing patients' conditions, these data are usually presented in a free-text format and are not standardized. They suggested employing language models to extract scale scores from the records.

More recently, several studies investigated the applicability of ML to predicting 30-day readmission after stroke[19], post-stroke activities of daily living (ADL)[20], stroke outcomes[21], and identifying patients with suspected stroke at the emergency department.[22]

More specifically, out of 74 features available in clinical records the authors[19] selected the top 20 of 6558 patients to train XGBoost in order to predict 30-day readmission. The authors also demonstrated that as few as 5 features were sufficient to accurately predict the readmission (AUC:0.76) while using 10 features further improved the model performance (AUC: 0.80). The features were selected using SHAP values.

The authors of the ADL study[20], employed logistic regression(LR), support vector machine, and random forest(RF) to predict the Barthel index status at discharge from rehabilitation, which was split into low the Barthel index(BI), medium BI and high BI categories. The models were trained on the clinical information comprising 17 features of 313 post-stroke patients. The authors compared the results of the ML models with the 3 single features (BI score on admission, instrumental activities of daily living scale(IADL), and Berg balance test(BBT). Among the three aforementioned features, BI on admission with no surprise was the best predictor of BI at discharge (AUC: 0.756), while among ML models both LR and RF performed comparably (AUC: 0.796).

The current pilot study has also employed ML to look at discharge outcome after stroke. However, there are several difference from previous studies in that it focuses on (a) predicting the outcome of hospital discharge with emphasis on death, (b) explanatory features are added at four stages to simulate the timeline of a stroke patient from admission to discharge, (c) the evolution of features' importance at each stage is investigated, (d) various combinations of stroke score systems are studied including contributions of other factors in predicting the outcome.

Methods

Patient selection and data extraction

This study was a retrospective study of patients treated in a comprehensive stroke unit for a period of 5 years (2015 – 2019). The study was approved by the ACT Health Human Research Ethics Committee (2021.LRE.00127, REGIS reference number 2021/ETH11170). Individuals were included if they had had an ischaemic stroke, either proven on MRI or CT scan or were diagnosed clinically to have had a stroke by an experienced stroke neurologist after review of their case notes. In addition, information about patient outcome destination was necessary for inclusion. Exclusion criteria comprised strokes presenting more than 24 h after onset, transient ischaemic attacks, or patients with a non-stroke diagnosis on discharge. Outcome discharge destinations were classified into three categories based on information as close to 90 days after stroke onset as possible (see below). Prediction of discharge destination was undertaken by using simple machine learning algorithms on limited, but easily available, demographic and clinical data in addition to the scores on standard stroke scales, namely the NIHSS and SSS.

Explanatory features extracted from the clinical notes of each patient included age, sex, ethnicity, history of previous stroke, premorbid modified Rankin Scale (mRS), admission NIHSS and SSS scores, type of stroke (ischaemic vs. haemorrhagic), known atrial fibrillation at the time of admission, known dyslipidemia, known hypertension, whether endovascular clot retrieval (ECR) was performed, intravenous thrombolysis was administered, or the patient was started on dual antiplatelet therapy, the NIHSS and SSS scores at 24 h, and the discharge destination.

In the 24-hour window, intravenous tPA (TT), intravascular clot retrieval (EVT) or dual antiplatelet therapy (DAPT) were performed according to standard clinical guidelines.

The outcome variable was the discharge destination which comprised three categories, namely 'home', 'rehabilitation' (which included both discharge to inpatient rehabilitation and/or nursing home) and 'death'. Patient numbers did not allow a distinction between discharge home without support and discharge home with support. Similarly, patients requiring further support in the form of rehabilitation or residential care were collapsed into one category ('rehabilitation') granted the relatively small numbers involved. The category 'death' included both patients who died in the hospital and patients who died

Table 1
The four assessment stages and relevant features available at each stage.

Stage	Included Features	Total num. of factors
1. pre-admission	age, sex, ethnicity, premorbid mRS, history of stroke	5
2. admission	as for Stage 1 plus atrial fibrillation, hypertension, dyslipidaemia, stroke type,initial NIHSS and SSS	11
3. treatment	as for Stage 2 plus whether dual antiplatelet agents were started, thrombolysis administered and/or ECR performed.	14
4. 24 h	as for Stage 3 plus 24-hour NIHSS and SSS	16

following discharge (but before 90 days from stroke onset).

Of note, the fact that patients' notes were reviewed retrospectively ensured that stroke mimics (in which the eventual diagnosis was not, in fact, stroke) and stroke chameleons (in which a genuine stroke presentation was initially misdiagnosed as something else) could be excluded for the purpose of this study.

It is also important to stress that the aim of this study was to try to determine outcomes at an early stage in the stroke journey, ideally before such complications as aspiration pneumonia, UTI, DVT and PE, NOMI and delirium would have had a chance to occur. This sort of clinical information was therefore not included in this study.

Prediction at different time points (stages)

In an effort to simulate information that would be available at different points along a patient's early stroke journey, the ability to predict outcome was determined at each of the four stages. Available information (features) gradually increased from one stage to the next. The relevant features at each stage are shown in Table 1.

Machine learning analysis

To determine which ML model might be most suited for use in this context, we examined three ML models, looking at the interplay of various explanatory features in predicting the discharge destination at each of the four stages. These models included *k*-nearest-neighbour (kNN) looking at different values of *k* and both Euclidean and City-block distances,[23] as well as two ensemble-based models: Adaptive Boosting (AdaBoost) and Bootstrap Aggregation (Bag).[24]

As an initial step to gain an overview of the data and assess whether there was clustering of the three discharge destinations, we visualised the data in two dimensions at each of the four stages using a 1-layer Autoencoder, a technique that slightly extends Principal Component Analysis by introducing nonlinearity in the principal component (latent) space.[25] We used a network with 2 latent neurons, and M inputs and outputs, where M was equal to the number of features available at each stage.

The three ML models were then assessed in terms of their overall performance at each of the four stages based on their accuracy in predicting discharge destination. A leave-one-out (LOO) cross-validation procedure was used to validate the performance of the models. Each set of tests involved removing a single patient while the remaining patients were used for training; therefore, this training / testing procedure was repeated N times, N being the total number of patients in each group. The LOO was selected to mitigate the impact of a small number of patients who died, unfortunately, this prevented us from computing p-values and confidence intervals.

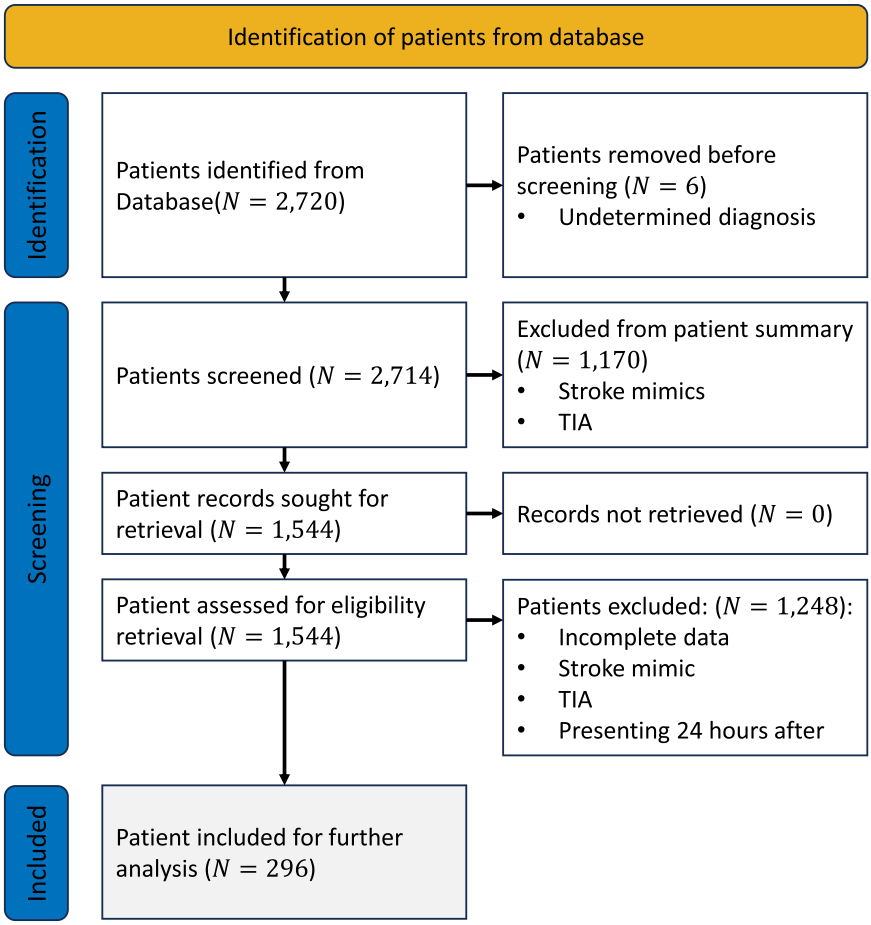


Fig. 1. PRISMA diagram to demonstrate selection of patients for ultimate inclusion in the study.

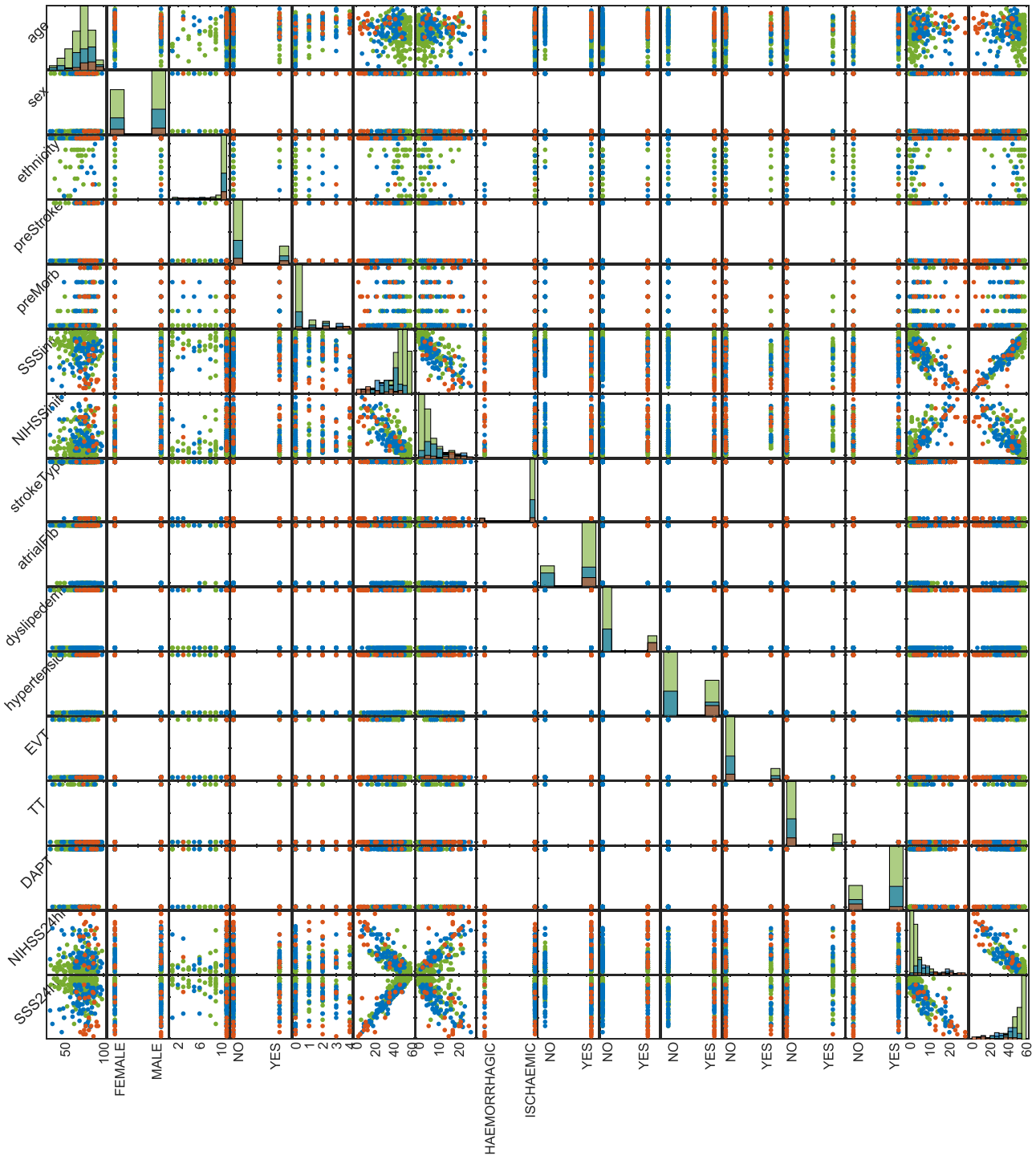


Fig. 2. Pairwise scatter plots for the 16 attributes present in the dataset. Both axes are equally labelled, although the labels on the x-axis are not shown due to size limitations, instead the values that corresponding attributes take, are shown. The charts on the main diagonal are class histograms for the corresponding attribute. Discharge destination is colour-coded as red for dead, blue for rehab, and green for home. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The per-class accuracy was computed as a ratio of correctly predicted outcomes divided by the total number of outcomes in a single class. The overall accuracy was defined as the sum of all correctly predicted outcomes divided by *N*. We also computed and presented the macro-averaged F1 score (F-macro), as a measure of model performance, as it is known to be robust to imbalanced datasets.

Overall patient numbers

The aim was to evaluate the models and determine whether their performance was consistent with both clinical experience and any information available in the literature. Consequently, we analyzed the contribution of individual attributes to predicting the outcome (discharge destination) at each of the four stages (Table 1) using ensemble-based feature importance and the Relief algorithm[25] that

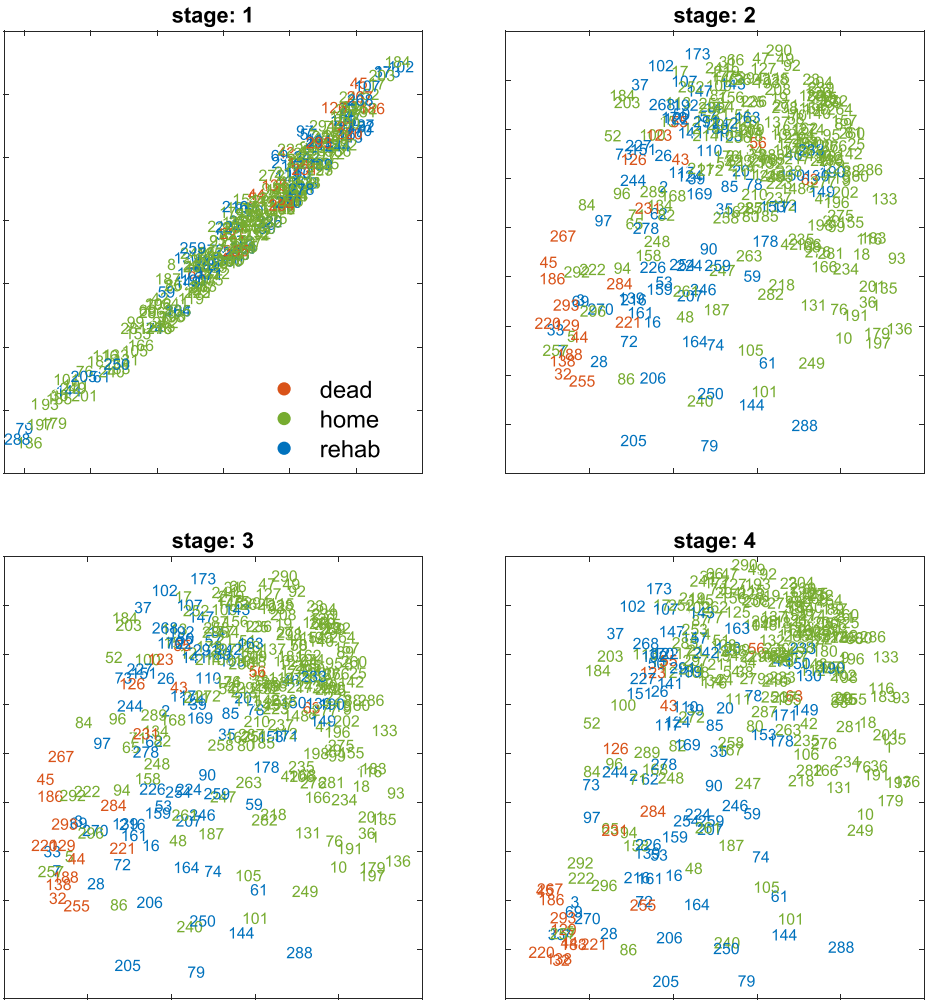


Fig. 3. Autoencoder-based feature-reduced space for outcomes at four stages. Each of the 296 patients is allocated one of three colours depending on the outcome destination at each of the four stages. The gradual increase in separation into three clusters of outcomes can be observed.

penalizes predictors giving different values to neighbors of the same class while rewarding predictors that give different values to neighbours of different classes.

In addition, to determine how much the change from initial to 24-hour stroke severity scores influenced the results, we examined the effect on overall precision of using only initial and only 24-hour scores. We were also interested in comparing the performance of the SSS and NIHSS scales. We looked at the change between initial and 24-hour scores as a function of discharge destination, as well as looking at the correlation between the scores of the two scales at both initial and 24-hour time points. To determine whether SSS or NIHSS might perform better for prediction purposes, we calculated the overall accuracy of each of the four models at each of the four stages limiting the models to the SSS or NIHSS score (but not both) at the initial and at 24-hour time points.

Results

The total number of patients admitted to the stroke unit over the period in question and the number of patients excluded is shown in Fig. 1. Many of the patient records were incomplete, particularly in relation to discharge destination and these patients were therefore excluded.

In summary, 296 patients (men: 175 (59.1%); women: 121 (40.9%); age: 31–98 years, median: 74 years) were included in the analysis. Of these, 200 (67.5%) went home, 76 (25.8%) went to rehabilitation, and 20 (6.8%) died. Of the 296 patients, 256 (87%) were Caucasian, 228 (78%) had dyslipidemia, and 183 (63%) had hypertension.

Fig. 2 presents pairwise scatter plots and per-class histograms for all 16 attributes (Table 1).

Visual inspection using Autoencoder

The results of the two-dimensional visualisation using Autoencoder are shown in Fig. 3. There was a significant improvement in the separation of patients from different discharge destinations in Stage 2 compared to Stage 1. Adding the three additional features of Stage 3 did not visually improve the separation of the patients compared to Stage 2, but the addition of the two additional features of Stage 4 improved the separation somewhat further. This visually observed clustering implied that ML should have the capacity to separate outcomes and, indeed, to predict them.

Overall accuracy of machine learning models at each of the four stages

Accuracies and F-scores of the individual models at each of the four stages are shown in Table 2. The error did not vary much with the number of neighbours when kNN with Euclidean distance was used, although classification error increased slightly as the number increased. The smallest error occurred at 12 and 14 nearest neighbours, so $k = 12$ was selected as representative. For City-block distance, $k = 2$ was selected as representative.

The overall accuracy of predicting discharge destination by all

Table 2
Overall and per-class accuracies. The numbers in bold font highlight the highest overall accuracies of the three models while the numbers shown in italic font highlight the best per-class prediction accuracies/F-score.

Stage 1	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	**75.9**	0.600	40.0	88.9	51.3
kNN (k = 2, Euclidean)	68.6	—	0.0	95.0	17.1
Bag	68.5	—	0.0	93.5	21.1
AdaBoost	67.1	—	0.0	93.2	18.4
Stage 2	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	**79.3**	0.678	50.0	88.9	59.2
kNN (k = 2, Euclidean)	77.4	**0.681**	55.0	89.5	51.3
Bag	73.9	0.639	55.0	84.4	47.4
AdaBoost	69.8	0.641	85.0	80.4	38.2
Stage 3	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	**78.6**	0.648	40.0	89.4	60.5
kNN (k = 2, Euclidean)	77.7	**0.684**	55.0	90.0	51.3
Bag	71.5	0.631	55.0	83.4	44.7
AdaBoost	66.2	0.619	85.0	80.5	31.6
Stage 4	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	**81.7**	0.716	45.0	89.4	71.1
kNN (k = 2, Euclidean)	77.7	0.668	40.0	87.0	63.2
Bag	74.2	0.647	50.0	85.4	51.3
AdaBoost	77.3	**0.727**	90.0	83.4	57.6

Table 3
Accuracies attained by the machine learning models using only initial SSS and initial NIHSS, only 24-hour SSS and 24-hour NIHSS scores, or all four scores. The numbers in bold font highlight the best scores across the three ML models while the numbers in italic font highlight the highest per-class accuracy for predicting death.

initSSS / initNIHSS	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	82.0	0.689	45.0	92.0	65.8
kNN (k = 2, Euclidean)	73.9	0.608	40.0	86.9	48.7
Bag	65.4	0.538	40.0	78.9	36.8
AdaBoost	67.8	0.585	45.0	82.9	34.2
24hSSS / 24hNIHSS	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	**80.6**	**0.700**	55.0	87.4	69.7
kNN (k = 2, Euclidean)	75.2	0.656	30.0	85.4	60.5
Bag	69.5	0.556	40.0	81.9	44.7
AdaBoost	72.8	0.599	40.0	84.4	51.3
InitSSS / InitNIHSS / 24hSSS / 24hNIHSS	overall		accuracy (%) per class		
	accuracy (%)	F-macro	dead	home	rehab
kNN (k = 2, City-block)	**80.3**	**0.656**	45.0	91.5	60.5
kNN (k = 2, Euclidean)	75.9	0.635	30.0	86.4	60.5
Bag	74.2	0.588	30.0	85.9	55.3
AdaBoost	70.5	0.545	30.0	84.4	44.7

models increased from Stage 1 to Stage 2. However, adding additional predictors (i.e. types of intervention) did not improve the performance further. kNN with City-block distance and 2 nearest neighbours outperformed kNN with Euclidean distance and 12 nearest neighbours. In particular, the accuracy of the former model at Stage 2 was 79.3%, increasing to 81.7% at Stage 4 while, for the latter model, the accuracy only reached 77.7%. The maximum accuracies using Bag and AdaBoost at Stage 4 were 74.2% and 77.3%, respectively.

Table 3 presents the accuracies resulting from using only the stroke scores at different time points, either initial SSS and NIHSS, 24-hour SSS and NIHSS, or all four. kNN/City-block achieved the best overall accuracies of 82.0% using only the initial scores and 80.6% when only 24-hour scores were used; there was no benefit to using all four scores. The maximum accuracies of Bag and AdaBoost were 74.2% and 72.8%, respectively. Again, there was little improvement conferred by including initial scores as well as 24-hour scores.

Accuracy of predicting specific outcomes

The ability of these models to predict outcome destination is shown in Table 2. The accuracy of predicting death using either the kNN model or Bag at Stage 2 was 50% – 55%, while at Stage 4 it was 40% – 50%, suggesting vulnerability to noise in the data. The accuracy using AdaBoost, however, was 85% at Stage 2 and 90% at Stage 4, suggesting that it was more robust and less affected by noise. In terms of predicting discharge home, all three models were similar with accuracies of 80% – 90% at Stage 2 and Stage 4. In predicting rehabilitation outcome, kNN performed best at all stages. Only AdaBoost demonstrated increased per-class accuracy for all three classes. Overall performance using only initial or only 24-hour stroke scores was reduced using all models to 65% – 80% with predicting death dropping to 30% – 55% (Table 3).

In summary, kNN ($k = 2$, City-block) had the best overall accuracy of 79.3% at Stage 2 and 81.7% at Stage 4 when all variables were included. Looking at individual outcomes, all models were able to predict discharge home with over 80% accuracy at all stages. Looking at death, AdaBoost consistently performed best with accuracies of 80% – 90% from Stage 2 onwards (Table 2). Rehabilitation was the hardest to predict, with the highest accuracies being achieved by kNN/City-block of 59.2% and 71.1% at Stages 2 and 4, respectively. By comparison, AdaBoost achieved 38.2% and 57.9%, respectively.

Contribution of individual feature to the prediction of outcome

The importance of each factor in relation to predicting discharge destination is shown in Fig. 4. Given that the LOO strategy was employed, for every model run, feature importance was estimated, resulting in 296 importance vectors for each stage. The brackets in Fig. 4 represent the standard deviation of each component in the importance vector.

The results of the Relief algorithm are shown in yellow: at Stage 2, the most important factors were initial SSS, pre-morbid mRS score and having had a previous stroke. At Stage 4, 24-hour SSS, 24-hour NIHSS, Initial SSS, pre-morbid mRS, and previous stroke were the variables that contributed most to the prediction of final outcome.

Using Bag (shown in red), the initial SSS score, premorbid mRS, and age were the highest contributors in Stage 2, while 24-hour SSS, 24-hour NIHSS and Initial SSS were the highest contributors in Stage 4. Using AdaBoost (shown in blue), both 24-hour scores were the most important predictors at Stage 4, although the SSS demonstrated higher importance than the NIHSS. Interestingly, dyslipidemia and hypertension were important to AdaBoost, but not to other models.

Comparison of SSS and NIHSS

Comparison of the initial and 24-hour scores and the comparison of the SSS and NIHSS scores are shown in Fig. 5. There was visual clustering by discharge destination, but not all patients fit into the clusters: Some of the patients who died were outliers, that is, their scores apparently improved over 24 h and yet they still died. Interestingly, the correlation between initial and 24-hour SSS scores was higher (0.93) than the corresponding correlation of NIHSS scores (0.82). This implies that the 24-hour SSS score provided little additional predictive information over and above the initial SSS score. Furthermore, there was a higher correlation between the two 24-hour scores than between the

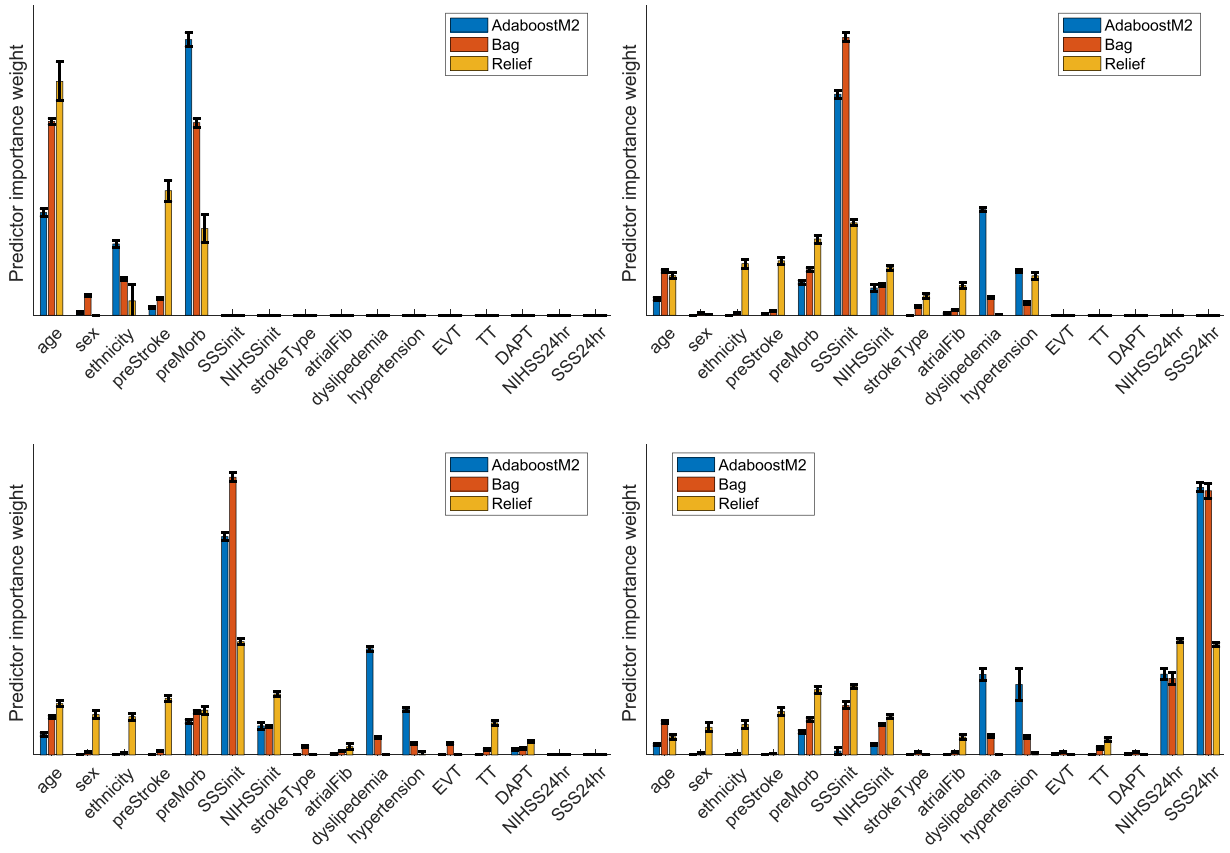


Fig. 4. Relative importance of factors in outcome prediction at each stage (Stages 1 to 4, from top to bottom). DAPT: dual anti-platelet therapy, EVT: endovascular therapy, Fib: fibrillation, TT: thrombolytic therapy.

initial scores, indicating that the two scores had a tendency to converge 24 h after the onset of the stroke.

This raised the question of whether one or the other scoring system might outperform the other. The effects of using either SSS or NIHSS scores alone, or in various combinations, are shown for each stage in Table 4. Interestingly, the highest predictive accuracy was obtained when the initial and 24-hour scores were derived from different scoring systems. Specifically, at Stage 4, a combination of initial SSS and 24-hour NIHSS scores resulted in the best overall performance using 2 out of 4 models, and in the remaining two, the accuracy was not too far behind the best accuracies. In terms of the prediction accuracy of the death outcome, the same score combinations were equal to or better than the accuracies of other score combinations attained by all models.

Discussion

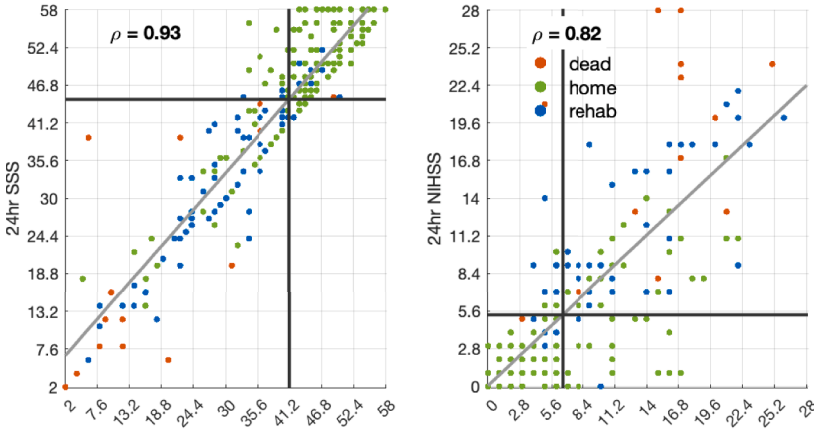
The process of progressively adding the available information in each of the four stages demonstrated a stepwise improvement in the ability of the three models to predict the outcomes. Of note, reasonable predictions were possible even at the time of arrival in the hospital, the pre-morbid mRS and initial SSS/NIHSS scores being as predictive as the scores at 24 h. Knowledge of which treatments were offered at stage 3 appeared to contribute relatively little to the models' ability to predict outcomes. There may have been many reasons for this lack of effect but the small numbers in the two intervention groups were likely to be responsible.

Three ML algorithms were compared to each other in terms of their ability to analyze the data. Though there was no single algorithm that could consistently predict all three outcomes better than the others, kNN with City-block distance was able to predict 89.9% of the patients that would go home at Stage 2, while Bag and AdaBoost correctly predicted 80.4% and 84.4% of patients, respectively. It was harder for the models to predict death because of the small numbers involved. However, AdaBoost significantly outperformed the other models, correctly predicting death in 85% (Table 2).

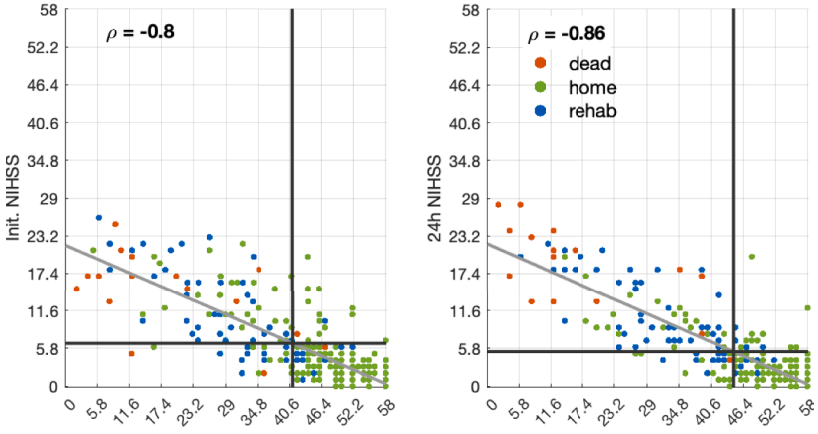
Looking at the prediction of stroke outcome after the intervention had occurred (i.e. at Stage 4), kNN ($k = 2$, City-block) had the highest overall accuracy (81.7%). When using kNN algorithms, it is possible that the use of small numbers of neighbours might result in overfitting. Consistent with this, increasing the number of neighbours and using different distance functions resulted in decreased overall accuracy: an example using 12 neighbours with Euclidean distance is shown in Table 2. AdaBoost, on the other hand, is recognised to be more resistant to overfitting than many other ML algorithms: applied to this data set, it was able to predict an outcome of death with 85% accuracy at Stage 3 and 90% accuracy at Stage 4, significantly outperforming the other models. Of note, dyslipidaemia and hypertension were important features when using AdaBoost as opposed to the other models. While these are well-known risk factors for ischaemic stroke, their inclusion could contribute to the higher predictive accuracy of AdaBoost. Overall, the data suggest that AdaBoost might be the most appropriate model to use to predict discharge outcomes, but this requires further confirmation (see below).

It is also worth noting that AdaBoost ranked ethnicity as the third most important factor at stage 1, with the importance value comparable to age. In total, 87% of the selected patients were White/Caucasian. Unfortunately, we did not have access to the socioeconomic background of the patients to further investigate its impact on discharge outcomes.

It should be noted that restricting the models to using only the SSS and NIHSS scores resulted in accuracies at each of the stages that were similar to those generated when the additional features were used. This demonstrates the high predictive power of stroke scores. However,



(a) Scatter plot of initial and 24-hour scores for SSS (left) and NIHSS (right)



(b) Scatter plot for initial (left) and 24-hour (right) SSS and NIHSS scores

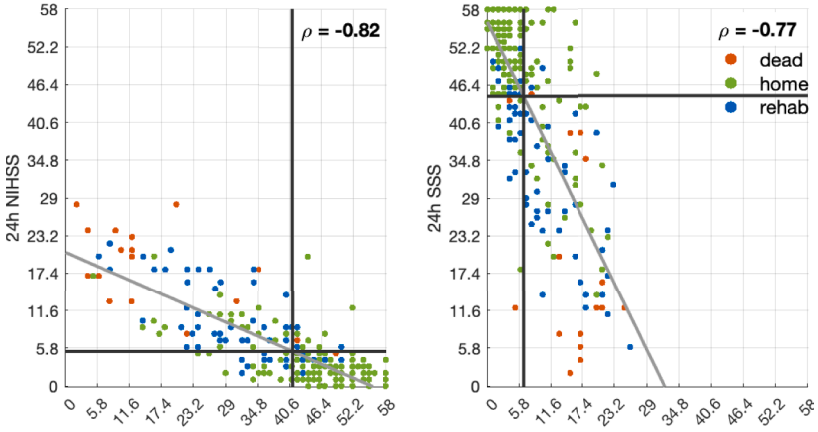


Fig. 5. Comparison of NIHSS and SSS scores at different time points. Outcome (discharge destination) is colour-coded as in Fig. 2.

looking at per-class accuracies for predicting death and rehabilitation, models using all available predictors performed better. In general, the models predicted home destination better than death or rehabilitation, but this is probably related to the fact that most (67.5%) patients went home, while only 6.8% died, and only 25.8% went to rehabilitation.

Looking at the factors that contributed most to prediction at Stages 1 and 2, premorbid mRS, initial SSS and initial NIHSS scores were, unsurprisingly, as important as age, previous history of stroke, hypertension, AF, and dyslipidaemia. In fact, the Relief algorithm found that premorbid mRS was important at all four stages. Broderick et al. discuss the strengths and limitations of mRS.[26] In the case of the two ensemble-based methods, though, premorbid mRS was less important, being overtaken by the initial SSS score at Stage 2. Previous studies have stressed the importance of variables like age, sex, previous stroke, hypertension, NIHSS, SSS, and mRS scores.[26–32] This study suggested that, at Stages 1 and 2 when a decision whether to intervene or not must be made, the initial SSS score and the premorbid mRS score were almost universally more important than any of the other variables, though AdaBoost identified dyslipidaemia as an important factor. Interestingly, the initial NIHSS score appeared to be a weaker contributor than the initial SSS score. SSS and NIHSS assess similar items so it is not surprising that they were highly correlated with each other (Fig. 5). The scores are, however, slightly different and this resulted in slightly different contributions to the ability to predict outcomes. Table 4 shows that the highest accuracies were obtained when initial and 24-hour scores were derived from different scoring systems: the combination of

Table 4
Accuracies (%) attained by the machine learning models using various combinations of initial and 24-hour SSS and NIHSS scores. The main values represent the overall accuracy, while the values in parentheses represent the per-class accuracy looking at death as an outcome. The numbers in bold font highlight the best scores across all four combinations of stroke scores at Stages 2 and 4 while the numbers in italic font highlight the highest per-class accuracy for death.

kNN/City-block	Stage 1	Stage 2	Stage 3	Stage 4
All scores	75.9 (40.0)	79.3(50.0)	78.6(40.0)	81.7(45.0)
initSSS/24hSSS		**80.7**(55.0)	78.6(40.0)	81.7(60.0)
initSSS/24NIHSS				**85.4**(60.0)
initNIHSS/24hSSS		79.3(50.0)	78.6(40.0)	82.0(55.0)
initNIHSS/24hNIHSS				83.7(55.0)
kNN/City-block	Stage 1	Stage 2	Stage 3	Stage 4
All scores	68.6 (0.0)	**77.4**(55.0)	77.7(55.0)	77.7(40.0)
initSSS/24hSSS		66.6(5.0)	66.6(45.0)	68.9(25.0)
initSSS/24NIHSS				76.7(50.0)
initNIHSS/24hSSS		75.0(45.0)	74.7(45.0)	**77.0**(50.0)
initNIHSS/24hNIHSS				76.7(50.0)
kNN/City-block	Stage 1	Stage 2	Stage 3	Stage 4
All scores	75.9 (40.0)	72.9(55.0)	71.5(55.0)	74.2(50.0)
initSSS/24hSSS		67.1(50.0)	71.5(50.0)	72.2(70.0)
initSSS/24NIHSS				75.9(55.0)
initNIHSS/24hSSS		70.2(55.0)	71.5(55.0)	75.9(55.0)
initNIHSS/24hNIHSS				73.9(50.0)
kNN/City-block	Stage 1	Stage 2	Stage 3	Stage 4
All scores	75.9 (40.0)	**69.8**(85.0)	68.2(85.0)	**77.3**(90.0)
initSSS/24hSSS		68.5(65.0)	61.5(85.0)	73.6(70.0)
initSSS/24NIHSS				75.9(95.0)
initNIHSS/24hSSS		68.1(85.0)	68.1(85.0)	73.6(70.0)
initNIHSS/24hNIHSS				72.5(95.0)

initial SSS and 24-hour NIHSS scores generated the highest overall accuracy using kNN/City-block and Bag, and a better performance when predicting death in all models. Hence mixing the scores may provide the optimum predictive power going forward.

An important question is whether ML adds anything to existing clinical scores. One published clinical score is the Stroke Prognostication using Age and NIHSS score (SPAN-100)[33], designed to predict the clinical outcome of thrombolysed patients. To evaluate this, the accuracies of the predictions of the ML models in the current study were compared with those of the SPAN-100 scores.[33] To align with the results of SPAN-100, we classified 'discharge home' as a favourable outcome and collapsed 'rehabilitation' and 'death' as unfavourable outcomes. Using our data, SPAN-100 correctly predicted a favourable outcome with 69.4% accuracy and an unfavourable outcome with 61.1% accuracy. By way of a fair comparison, these scores were compared to the outcome of the ML models at Stage 2 (i.e. before intervention). A favourable outcome was predicted by the ML models 78.6% – 82.3% of the time while an unfavourable outcome was accurately predicted 58.7% – 73.1% of the time. This demonstrates the ability of ML to enhance, and therefore outperform, existing clinical tools.

Limitations

This study was a pilot study from a single tertiary care hospital and was only able to assess a very limited sample size of 296 patients because of incomplete information available in the patients' notes. Because of the relatively small numbers involved, it was necessary to collapse outcome groups which would, ideally, have been left separate. In addition, the majority of our patients were white, meaning that the possibility that ethnic background might have influenced outcome destination could not be assessed. Similarly, socioeconomic factors might have contributed to the outcome but these were not studied here as our medical records did not provide adequate information.

Premorbid factors such as dyslipidaemia and hypertension were included for the purposes of exploration without proposing any specific mechanism whereby they might have actually influenced the ultimate discharge destination. The fact that they clearly appeared to influence outcome deserves further exploration and explanation.

Future prospective studies should provide more definitive information, and this study strongly suggests that such studies would be worthwhile. These future studies should investigate additional clinical information, such as socioeconomic factors and the existence of diabetes mellitus before stroke onset, in addition to initial clinical assessments such as blood pressure, temperature and blood sugar, and the results of initial blood tests at the time of presentation.

Finally, it is important to note that this study only investigated three possible ML models. Future studies could broaden the scope with a view to determining which model provides the most reliable results and would therefore be most helpful to clinicians in clinical use.

The way forward

This study demonstrated that ML models have the potential to improve predictive accuracy in the initial stages of acute stroke management. Different ML algorithms performed better with respect to different outcomes, but in general AdaBoost appeared to perform best in this study. However, in reality, no single ML algorithm consistently outperformed the others. Further clarification would be obtained from a larger prospective study that could also explore the optimal use of NIHSS and SSS scores. The findings of radiological investigations and the clinical location of the stroke could be incorporated into future studies. Ultimately, a randomized controlled trial will be required to determine whether the incorporation of ML algorithms into predictive models can usefully influence clinical decision-making, both in terms of patient outcome and cost-effectiveness.

Conclusion

This was a pilot study that evaluated a small cohort of stroke patients, looking at a limited number of factors and ML models. Considering the ability of the included ML models to predict outcome on the basis of such a small subset of factors, a second prospective phase of this study is proposed, looking at a larger cohort with a larger number of variables aiming to address some of the above limitations.

All models demonstrated comparable performance, while AdaBoost appeared to be the most robust model in predicting death. Importantly, the ML models outperformed clinical tools, such as the SPAN-100 score. Premorbid mRS and SSS scores had the highest feature importance in predicting discharge outcomes, while dyslipidemia and hypertension were important factors in predicting death. Overall, we conclude that the ML models have the capacity to be useful to clinicians in predicting discharge outcomes at an early stage in the management of patients presenting with acute stroke, and this information is potentially useful in influencing subsequent clinical management decisions. Finally, an analysis of the various individual components of the scoring systems has been performed showing that some score combinations perform better than others and even better when all scores are used.

Funding

This work was partially funded by Lensky Analytics and by Our Health in Our Hands (OHIOH), a strategic initiative of the Australian National University (ANU), whose objective is to transform healthcare by developing new personalized health technologies and solutions in collaboration with patients, clinicians, and health care providers. We gratefully acknowledge the funding from the ANU School of Computing for Robin Vlieger's PhD studies.

CRediT authorship contribution statement

Artem Lensky: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Christian Lueck:** Conceptualization, Validation, Formal analysis, Investigation, Supervision, Methodology, Writing – original draft, Writing – review & editing. **Hanna Suominen:** Project administration, Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Brett Jones:** Conceptualization, Data curation, Methodology. **Robin Vlieger:** Investigation, Conceptualization, Methodology. **Tina Ahluwalia:** Investigation, Data curation, Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Computational methods and development of machine learning methods used in analysing of the data presented in this study was performed with financial support of Lensky Analytics. Initially this work begun with the financial support provided by the ANU OHIOH initiative. Hanna Suominen and Christian Lueck belong to the Chief Investigators of OHIOH. Artem Lensky is the CEO of Lensky Analytics

References

1. Deloitte access economics. the economic impact of stroke in australia. https://www2.deloitte.com/content/dam/Deloitte/au/Documents/Economics/deloitte-au-dae-economic-impact-stroke-report-061120.pdf, Accessed 2023-02-14; 2020.
2. The no postcode untouched: stroke in Australia 2020 report. https://strokefoundation.org.au/media/juuba3qm/no-postcode-untouched-30-october-final-report.pdf, Accessed 2023-02-14; 2020.
3. Zhelev Z, Walker G, Henschke N, Fridhandler J, Yip S. Prehospital stroke scales as screening tools for early identification of stroke and transient ischemic attack. *Emergencias*. 2021;33(4):312–314.
4. Antipova D, Eadie L, Macaden AS, Wilson P. Diagnostic value of transcranial ultrasonography for selecting subjects with large vessel occlusion: a systematic review. *Ultrasound J*. 2019;11(1):29.
5. Lima FO, Silva GS, Furie KL, et al. Field assessment stroke triage for emergency destination: a simple and accurate prehospital scale to detect large vessel occlusion strokes. *Stroke*. 2016;47(8):1997–2002.
6. Luvizutto GJ, Monteiro TA, Braga G, Pontes-Neto OM, de Lima Resende LA, Bazan R. Validation of the scandinavian stroke scale in a multicultural population in brazil. *Cerebrovasc Dis Extra*. 2012;2(1):121–126.
7. Jumaa MA, Castonguay AC, Salahuddin H, et al. Long-term implementation of a prehospital severity scale for EMS triage of acute stroke: a real-world experience. *J Neurointerv Surg*. 2020;12(1):19–24.
8. Torres-Aguila NP, Carrera C, o E, et al. Clinical variables and genetic risk factors associated with the acute outcome of ischemic stroke: a systematic review. *J Stroke*. 2019;21(3):276–289.
9. Emdad FB, Tian S, Nandy E, Hanna K, He Z. Towards interpretable multimodal predictive models for early mortality prediction of hemorrhagic stroke patients. *AMIA Jt Summits Transl Sci Proc*. 2023;2023:128–137.
10. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artif Intell Med*. 2023;143:102616. https://doi.org/10.1016/j.artmed.2023.102616.
11. Rana S, Luo W, Tran T, et al. Application of machine learning techniques to identify data reliability and factors affecting outcome after stroke using electronic administrative records. *Front Neurol*. 2021;12:670379.
12. Lella L, Gentile L, Pristipino C, Toni D. *Predictive clustering learning algorithms for stroke patients discharge planning. HEALTHINF 2021 - 14th International Conference on Health Informatics*. SciTePress; 2021:296–303.
13. Counsell C, Dennis M, McDowall M, Warlow C. Predicting outcome after acute and subacute stroke: development and validation of new prognostic models. *Stroke*. 2002;33(4):1041–1047.
14. Teale EA, Forster A, Munyombwe T, Young JB. A systematic review of case-mix adjustment models for stroke. *Clin Rehabil*. 2012;26(9):771–786.
15. nig IR, Ziegler A, Bluhmki E, et al. Predicting long-term outcome after acute ischemic stroke: a simple index works in patients from controlled clinical trials. *Stroke*. 2008;39(6):1821–1826.
16. Weimar C, nig IR, Kraywinkel K, Ziegler A, Diener HC. Age and national institutes of health stroke scale score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia: development and external validation of prognostic models. *Stroke*. 2004;35(1):158–162.
17. Chen D, Zhu Y, Wang Y, et al. A new clinical score to predict the possibility of stroke patients receiving intravenous thrombolysis. *J Stroke Cerebrovasc Dis*. 2023;32(4):107037.
18. Yang L, Huang X, Wang J, et al. Identifying stroke-related quantified evidence from electronic health records in real-world studies. *Artif Intell Med*. 2023;140:102552. https://doi.org/10.1016/j.artmed.2023.102552.
19. Lv J, Zhang M, Fu Y, et al. An interpretable machine learning approach for predicting 30-day readmission after stroke. *Int J Med Inform*. 2023;174:105050.
20. Lin WY, Chen CH, Tseng YJ, et al. Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. *Int J Med Inform*. 2018;111:159–164.
21. Lin CH, Hsu KC, Johnson KR, Luby M, Fann YC. Applying density-based outlier identifications using multiple datasets for validation of stroke clinical outcomes. *Int J Med Inform*. 2019;132:103988.
22. Sung SF, Hung LC, Hu YH. Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning. *Int J Med Inform*. 2021;152:104505.
23. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;25(5):1189–1232.
24. Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RRelieff. *Mach Learn*. 2003;53:23–69.
25. Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big Data*. 2019;7(4):221–248.
26. Broderick JP, Adeoye O, Elm J. Evolution of the modified Rankin scale and its use in future stroke trials. *Stroke*. 2017;48(7):2007–2012.
27. Meyer MJ, Pereira S, McClure A, et al. A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation. *Disabil Rehabil*. 2015;37(15):1316–1323.
28. Ferriero G, Franchignoni F, Benevolo E, Ottonello M, Scocchi M, Xanthi M. The influence of comorbidities and complications on discharge function in stroke rehabilitation inpatients. *Eura Medicophys*. 2006;42(2):91–96.
29. Scrutinio D, Lanzillo B, Guida P, et al. Development and validation of a predictive model for functional outcome after stroke rehabilitation: the Maugeri model. *Stroke*. 2017;48(12):3308–3315.
30. Gialanella B, Santoro R, Ferlucci C. Predicting outcome after stroke: the role of basic activities of daily living predicting outcome after stroke. *Eur J Phys Rehabil Med*. 2013;49(5):629–637.
31. Askim T, Bernhardt J, Churilov L, Indredavik B. The scandinavian stroke scale is equally as good as the national institutes of health stroke scale in identifying 3-month outcome. *J Rehabil Med*. 2016;48(10):909–912.
32. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. *Stroke*. 2007;38(3):1091–1096.
33. Saposnik G, Guzik AK, Reeves M, Ovbiagele B, Johnston SC. Stroke prognostication using age and NIH stroke scale: SPAN-100. *Neurology*. 2013;80(1):21–28.