Evaluation of Machine Learning Methods to Stroke Outcome Prediction Using a Nationwide Disease Registry

Ching-Heng Lin , Kai-Cheng Hsu , Kory R. Johnson , Yang C. Fann , Chon-Haw Tsai , Yu Sun , Li-Ming Lien , Wei-Lun Chang , Po-Lin Chen , Cheng-Li Lin , Chung Y. Hsu , Taiwan Stroke Registry Investigators

PII: S0169-2607(19)31436-1
DOI: https://doi.org/10.1016/j.cmpb.2020.105381
Reference: COMM 105381

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 27 August 2019
Revised date: 31 December 2019
Accepted date: 31 January 2020

Please cite this article as: Ching-Heng Lin , Kai-Cheng Hsu , Kory R. Johnson , Yang C. Fann , Chon-Haw Tsai , Yu Sun , Li-Ming Lien , Wei-Lun Chang , Po-Lin Chen , Cheng-Li Lin , Chung Y. Hsu , Taiwan Stroke Registry Investigators, Evaluation of Machine Learning Methods to Stroke Outcome Prediction Using a Nationwide Disease Registry, *Computer Methods and Programs in Biomedicine* (2020), doi: https://doi.org/10.1016/j.cmpb.2020.105381

Highlights

- Using a nationwide prospective stroke registry to evaluate several machine learning approaches for prediction of stroke outcomes.
- Over two hundred clinical variables are screened to identify important features that predicts stroke outcome.
- The follow-up data is important which can further improve the predictive models' performance.
- Error analysis shows that most prediction errors come from more severe stroke patients.

Evaluation of Machine Learning Methods to Stroke Outcome Prediction Using a Nationwide Disease Registry

Ching-Heng Lin[1,2], Kai-Cheng Hsu[2,3], Kory R. Johnson[2], Yang C. Fann,[2*], Chon-Haw Tsai,[4], Yu Sun,[5], Li-Ming, Lien,[6,7], Wei-Lun Chang,[8], Po-Lin Chen,[9, 10], Cheng-Li Lin[11], Chung Y. Hsu[12] and Taiwan Stroke Registry Investigators[#]

[1] Center for Information Technology, National Institutes of Health, Bethesda, Maryland, United States.
[2]Bioinformatics Section, National Institute of Neurological Disorder and Stroke, National Institutes of Health, Bethesda, Maryland, United States.
[3]Department of Neurology, National Taiwan University Hospital, Taipei, Taiwan
[4] Division of Nephrology, China Medical University Hospital, Taichung, Taiwan
[5]Neurology, En Chu Kong Hospital, New Taipei City, Taiwan
[6]Department of Neurology, Shin Kong Wu-Ho-Su Memorial Hospital, Taipei, Taiwan
[7]Department of Neurology, College of Medicine, Taipei Medical University, Taipei, Taiwan
[8]Department of Neurology, Show Chwan Memorial Hospital, Changhua County, Taiwan
[9]Neurological Institute, Taichung Veterans General Hospital, Taichung, Taiwan
[10]Department of Neurology, School of Medicine, National Yang-Ming University, Taipei, Taiwan.
[11] Management Office for Health Data, China Medical University Hospital, Taichung, Taiwan
[12]Graduate Institute of Biomedical Sciences, China Medical University, Taichung, Taiwan.
[*]Corresponding author
[#]List in appendix A

Corresponding Author:
Yang C. Fann, Ph.D.
Director, Intramural IT & Bioinformatics Program
National Institutes of Neurological Disorders and Stroke
National Institutes of Health
9000 Rockville Pike, Bethesda, Maryland 20892
Tel: 301-451-5153, Fax: +301-480-3563,
E-mail: fann@ninds.nih.gov

Abstract

Introduction

Being able to predict functional outcomes after a stroke is highly desirable for clinicians. This allows clinicians to set reasonable goals with patients and relatives, and to reach shared after-care decisions for recovery or rehabilitation. The aim of this study was to apply various machine learning (ML) methods for 90-day stroke outcome predictions, using a nationwide disease registry.

Methods

This study used the Taiwan Stroke Registry (TSR) which has prospectively collected data from stroke patients since 2006. Three known ML models (support vector machine, random forest, and artificial neural network), and a hybrid artificial neural network were implemented and evaluated by 10-time repeated hold-out with 10-fold cross-validation.

Results

ML techniques present over 0.94 AUC in both ischemic and hemorrhagic stroke using preadmission and inpatient data. By adding follow-up data, the prediction ability improved to 0.97 AUC. We screened 206 clinical variables to identify 17 important features from the ischemic stroke dataset and 22 features from the hemorrhagic stroke dataset without losing much performance. Error analysis revealed that most prediction errors come from more severe stroke patients.

Conclusion

The study showed that ML techniques trained from large, cross-reginal registry datasets were able to predict functional outcome after stroke with high accuracy. The follow-up data is important which can further improve the predictive models' performance. With similar performances among different ML techniques, the algorithm's characteristics and performance on severe stroke patients will be the primary focus when we further develop inference models and artificial intelligence tools for potential medical.

Keywords: stroke outcome; machine learning; ischemic stroke; hemorrhagic stroke

1. Introduction

Stroke is the second leading cause of mortality in the world and the leading adult disability in developed countries [1, 2]. Many stroke survivors are left with various neurological deficits resulting in impaired quality of life of variable extent that has been a significant burden on patients, caregivers, and society [3]. More precise prediction of functional outcomes after a stroke may help clinicians in developing an appropriate long-term management plan. For example, plans based on better prediction of the extent of recovery with appropriate rehabilitative measures with patients' domestic condition taken into consideration for reaching shared decisions with patients and family members [4-6]. Much effort has been devoted to determining predictors of functional outcome after stroke [6-8]. Several medical communities have created scores that can predict the patient's functional outcome using data readily available at admission [9-11]. These scores use statistical analysis to identify the most relevant covariates from a set of pre-selected factors by domain experts. Recently, machine learning has become ubiquitous for solving complex problems in many scientific domains, especially in medical diagnosis or prognosis prediction [12, 13].

With Institutional Review Board approval, the Taiwan Stroke Registry (TSR) program began in August 2006 and prospectively collected stroke patients treatment information from 64 major hospitals in Taiwan [14]. The TSR database was the first national stroke database to assess the quality of stroke care. The data were systematically collected according to predetermined registry protocols. To ensure data reliability, the participating hospital neurologists and nurses, responsible for completion of registration materials were trained with TSR's standard operating procedure. The TSR mainly collects preadmission data, and inpatient elements including clinical care during hospitalization, in-hospital complications, stroke risk factors, laboratory results of blood tests, electrocardiography, computed tomography (CT) and magnetic resonance imaging (MRI) findings, medications during admission and discharge status. In addition, the TSR also collects follow-up information such as stroke outcome, patient's location and vascular events during one-year period (30, 90, 180 and 360 days respectively).

This study aims to evaluate the performance of supervised machine learning models using clinical features including 30 days follow-up data that can predicts 90-day stroke outcome which is known to be highly correlated to the future recovery and wellbeing of stroke patients. Our approach considers the problems of data preprocessing, feature selection, and prediction in

stroke and its subtype datasets. Prognostic training models were selected via stratified 10-fold cross validation and were assessed with holdout data ten times. To conduct this study, ethical clearance was approved from the Joint Institutional Review Board (JIRB) of CMUH102-REC l-086(CR-5) at China Medical University of Taiwan.

2. Materials and Methods

2.1 Data Source and Preprocessing

This study used the TSR records of stroke patients who had been documented in the TSR database between 2006 and 2018. The registry required each participating hospital to differentiate the stroke types using CT and/or MRI. All patients had signed informed consent documents and the identity of patients were scrambled to protect privacy.

We performed a series of extract-transform-load process to get 58,493 stroke patient records form the TSR. Figure 1 lists four data exclusion criteria applied for this study: (1) patients who died before discharge, (2) patients who has no 30-day follow-up information, (3) patients who is other type of stroke (not ischemic/heorragic brain stroke) and (4) patients who has illogical or contradicted assessments. This study used the 90-day modified Rankin Scale (mRS; 0 (no symptoms) - 6 (death)) as the measure of patient's outcome since it shows better clinimetric properties for assessing the impact of stroke treatments.

During data quality assurance, we found there are mismatches between discharge mRS and the patients' discharge Barthel index (BI; 100 (independent) – 0 (dependent)) total score. The mRS and Barthel index both are behavior assessment, and they showed a significant negative correlation and distribution overlapped [15, 16] that can be a reference for assessment validation. As Figure 2 shows, a patient with an mRS score zero (no symptoms) but has a BI total score of less than 20 (totally dependent) at the time of discharge. To eliminate the observations with illogical assessments, we applied an assessment validation during the data preprocessing flow to ensure data validity. The assessment validation consists of two processes, one is the clinical-logic validation, and the second is a non-linear regression method. In the clinical-logic validation, we formulated a set of logic rules based on the assessment principles and medical knowledge to validate the data. In non-linear regression, we applied the locally weighted scatterplot smoothing (LOWESS) algorithm [17] to remove illogical assessments. After the LOWESS process, we calculated the trimmed average standard deviation using the

standard deviation of discharge mRS values per BAR bin. The observations that fall outside one standard deviation of the trimmed average were considered as an outlier. After the validation process, we indicated 10,542 observations with invalidated assessments. After the application of the exclusion criteria, the whole dataset was separated into an ischemic stroke dataset with 35,798 cases, and a hemorrhagic stroke dataset with 4,495 cases. We further dichotomized the patient's 90-day outcome into good outcome (mRS≤2) and poor outcome (mRS ≥3) [18-20] (Figure 1). Two feature sets were created from the TSR, one contains preadmission and inpatient data (clinical feature set with 203 features), and the other set additionally contains follow-up information (whole feature set with 206 features).

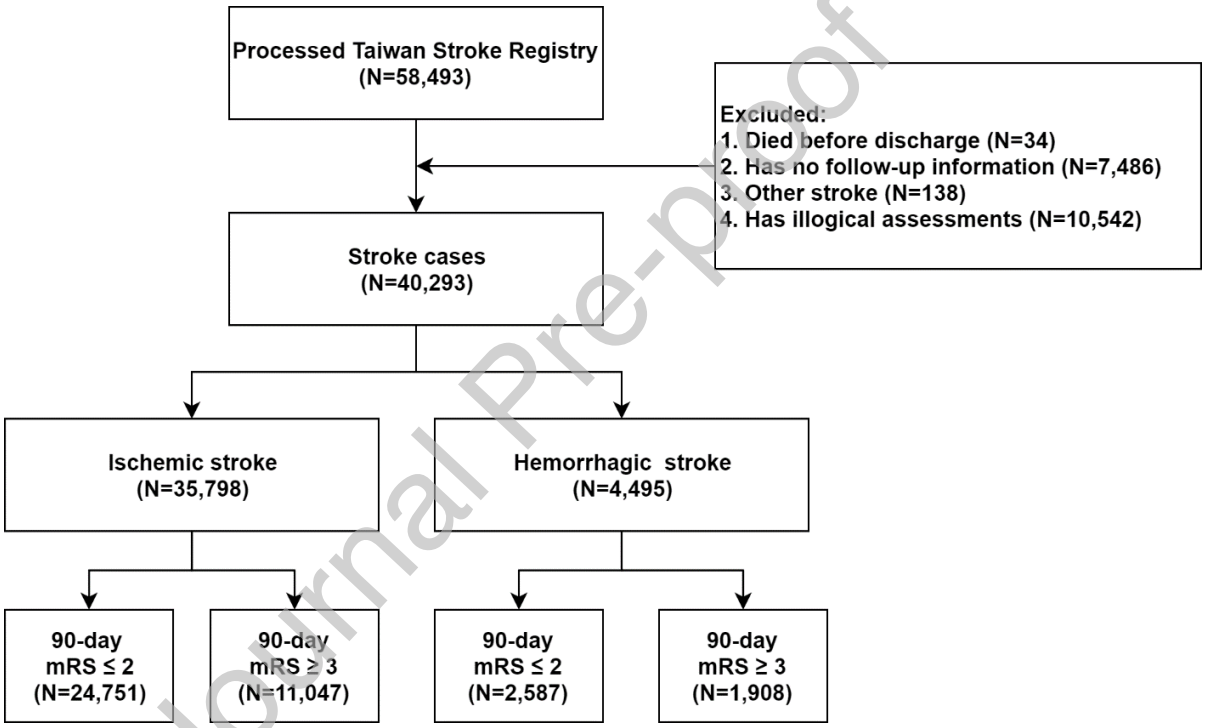


Figure1. Flowchart of patient recruitment. mRS indicates modified Rankin Scale.

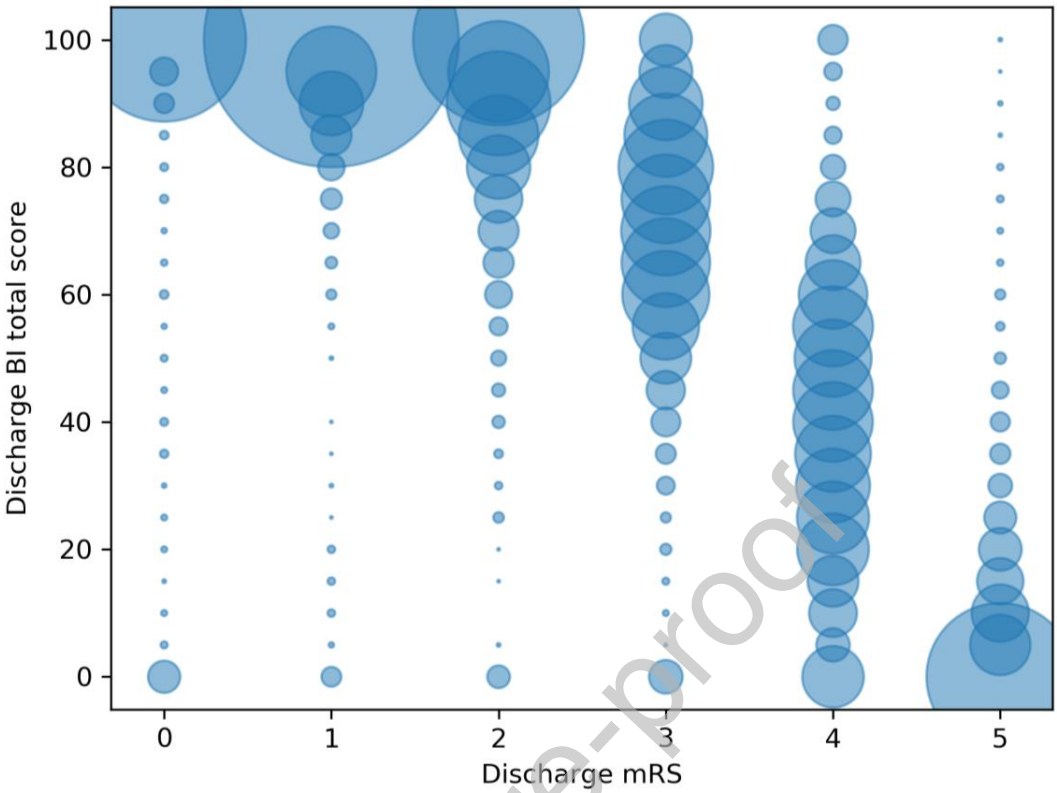


Figure 2. Bubble plot of discharged mRS degree and discharged Barthel index total score. The size of bubble indicates the size of population.

2.2 Machine Learning Models

In order to develop a prognostic model for the 90-day mRS outcome, this study assessed various supervised machine learning algorithms in terms of their ability to predict outcomes. High-performance machine learning algorithms such as support vector machine (SVM) [21], random forest (RF) [22], and artificial neural network (ANN) [23] were explored for comparison. We also design a hybrid artificial neural network (HANN) which parallelly combined dot product layers and fully connected layers to identify the pattern of various types of clinical data. To implement these machine learning algorithms for this study, we used the libraries of scikit-learn 0.19.2 [24] and Keras neural networks API [25]. The details of machine learning models can be found in appendix B.

2.3 Cross Validation

We designed a 10-time repeated hold-out with 10-fold cross-validation to assess the predictive capabilities of machine learning models. Figure 3 illustrates the details of the cross-validation process. For each stroke type, we used 70% of the data for model training and 30% of

the data for testing. In order to avoid training bias caused by the imbalance dataset, we applied down-sampling and shuffling on two stroke type training datasets, but the distribution of the testing dataset remained the same as the original dataset. In order to reduce the selection bias, the hold-out method was repeated 10 times. In each hold-out round, the training data were used for feature selection and best model selection with 10-fold cross-validation. The model with the highest accuracy was selected as the best model for further evaluation. Each prognostic model was tested by ischemic and hemorrhagic stroke hold-out datasets by 10 times with different feature sets. The performance measurements included accuracy:$\frac{TP+TN}{total\ population}$, precision: $\frac{TP}{TP+FP}$, recall: $\frac{TP}{TP+FN}$, F1-score:$\frac{2*precision*recall}{precision+recall}$ and receiver operating curve (ROC), where *TP*, *TN*, *FP* and *FN* denotes true positives, true negative, false positives and false negatives, respectively. Through this repeating process, we can achieve robust feature selection, and fairly estimate the performance of the machine learning models.

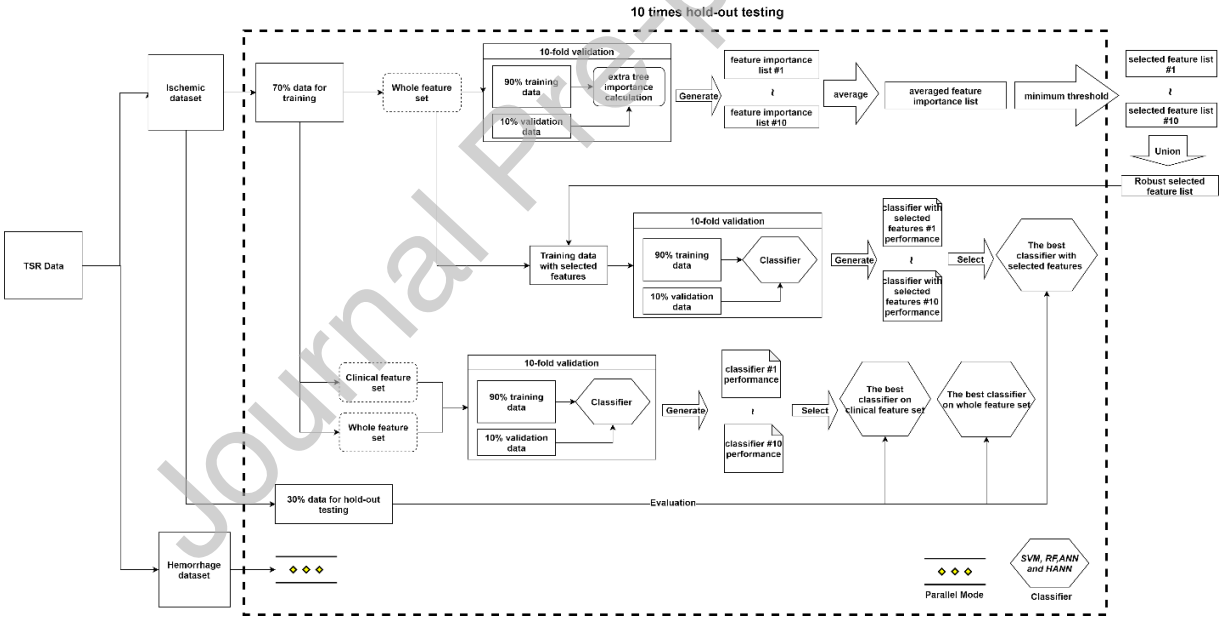


Figure 3. Flow chart of 10-time repeated hold-out with 10-fold cross-validation.

2.4 Feature Selection

Selection of relevant input features for outcome prediction is a common task in most machine learning modeling studies. The main idea is trying to identify the smallest possible set of input features that can still achieve good predictive performance. Furthermore, reducing the feature dimension can decrease the chance of overfitting, as well as eliminating uninformative

features, to let the classifier focus on informative variables. This study applied extremely randomized trees (extra-trees) algorithm [26] to perform the feature selection task on a whole feature set. In the extra-tree's algorithm, both feature and cut-point are randomly chosen while splitting a tree node at each iteration; therefore, the sequence of feature importance list can change. It is better to repeat the extra forest feature selection to get a robust selected feature list. At each hold-out round in our cross-validation, the feature's importance was calculated by extra-trees algorithm. The importance was defined by Gini impurity: $1 - \sum_{i=1}^{j} p_i^2$, where j is the number of classes, p_i is the fraction of items labeled with class i. After 10-fold rounds, we averaged the features importance and selected important features by minimum threshold. The minimum threshold for each iteration was defined as $threshold_{min} = \min(\sigma) + sd(\sigma)$, where σ is the list of feature importance in which the zeros have been removed [27]. After 10-times hold-out rounds, we took the union of ten selected features list to get the robust selected feature list.

3. Results

3.1 Experiment Data Description

The characteristics of the TSR experiment dataset are summarized by gender and shown in table 1. Male patients account for 61.96% of the total cases, and male patient average onset age is younger than female patients (65.4 vs. 69.71). In the type of stroke case, cerebral infarction stroke was more common in both females (79.5%) and males (79.7%). For the severity of stroke, 46.2% of female patients and 52.1% of male patients are classified as having a minor stroke ($1 \leq$ NIHSS ≤ 4). Most of the patients' Barthel index total score was between 80 and 100 (57.9% female and 70.0% male).

Table 1: Patient characteristics of Taiwan stroke registry experiment dataset

	Female	Male	P value	SMD
Case number (%)	15,328 (38.04)	24,965 (61.95)		
Onset age (Mean ± SD)	69.71 ± 12.62	65.40 ± 12.64	<0.001	0.342
Diagnosis			<0.001	0.112
Ischemic stroke				
Infarct (%)	12,182 (79.5)	19,901 (79.7)		
Transient (%)	1,400 (9.1)	2,315 (9.3)		
Hemorrhagic stroke				
Intracerebral (%)	1,413 (9.2)	2,537 (10.2)		
Subarachnoid (%)	333 (2.2)	212 (0.8)		
Discharge mRS[*]			<0.001	0.253
Good outcome (%)	7,979 (52.1)	16,081 (64.4)		
Poor outcome (%)	7,349 (47.9)	8,884 (35.6)		
Discharge total Barthel index[#]			<0.001	0.289
Independent (%)	8,871 (57.9)	17,466 (70.0)		
Minimally dependent (%)	1,828 (11.9)	2,658 (10.6)		
Partially dependent (%)	1,524 (9.9)	2,048 (8.2)		
Very dependent (%)	771 (5.0)	832 (3.3)		
Totally dependent (%)	2,334 (15.2)	1,961 (7.9)		
Discharge total NIHSS[$]			<0.001	0.207
No stroke symptoms (%)	3,349 (21.8)	5,896 (23.6)		
Minor stroke (%)	7,087 (46.2)	13,001 (52.1)		
Moderate stroke (%)	3,580 (23.4)	5,006 (20.1)		
Moderate to severe stroke (%)	562 (3.7)	473 (1.9)		
Severe stroke (%)	750 (4.9)	589 (2.4)		

SMD: standardized mean difference

[*] Good outcome: mRS ≤ 2; Poor outcome: mRS ≥ 3 [28].

[#] Independent: 80~100; Minimally dependent: 60~79; Partially dependent: 40~59; Very dependent: 20~39; Totally dependent: < 20 [29].

[$] No stroke symptoms: 0; Minor stroke: 1~4; Moderate stroke: 5~15; Moderate to severe stroke: 16~20; Severe stroke: 21~42 [30].

3.2 Important Features

We applied extra-trees algorithm with 10-times repeated hold-out with 10-fold cross-validation to select robust important features. We selected 17 features for the ischemic stroke dataset and 22 features for the hemorrhagic stroke dataset from the 206 various features in whole feature set. The selected features are displayed as single heatmap and sorted by its importance in Figure 4. The numbers indicate how many times the feature passes the minimum threshold during cross-validation. For both stroke types, most of selected features are discharge NIHSS assessment items, discharge Barthel index, and the 30-day mRS degree which is the most important feature for 90-day mRS prediction.

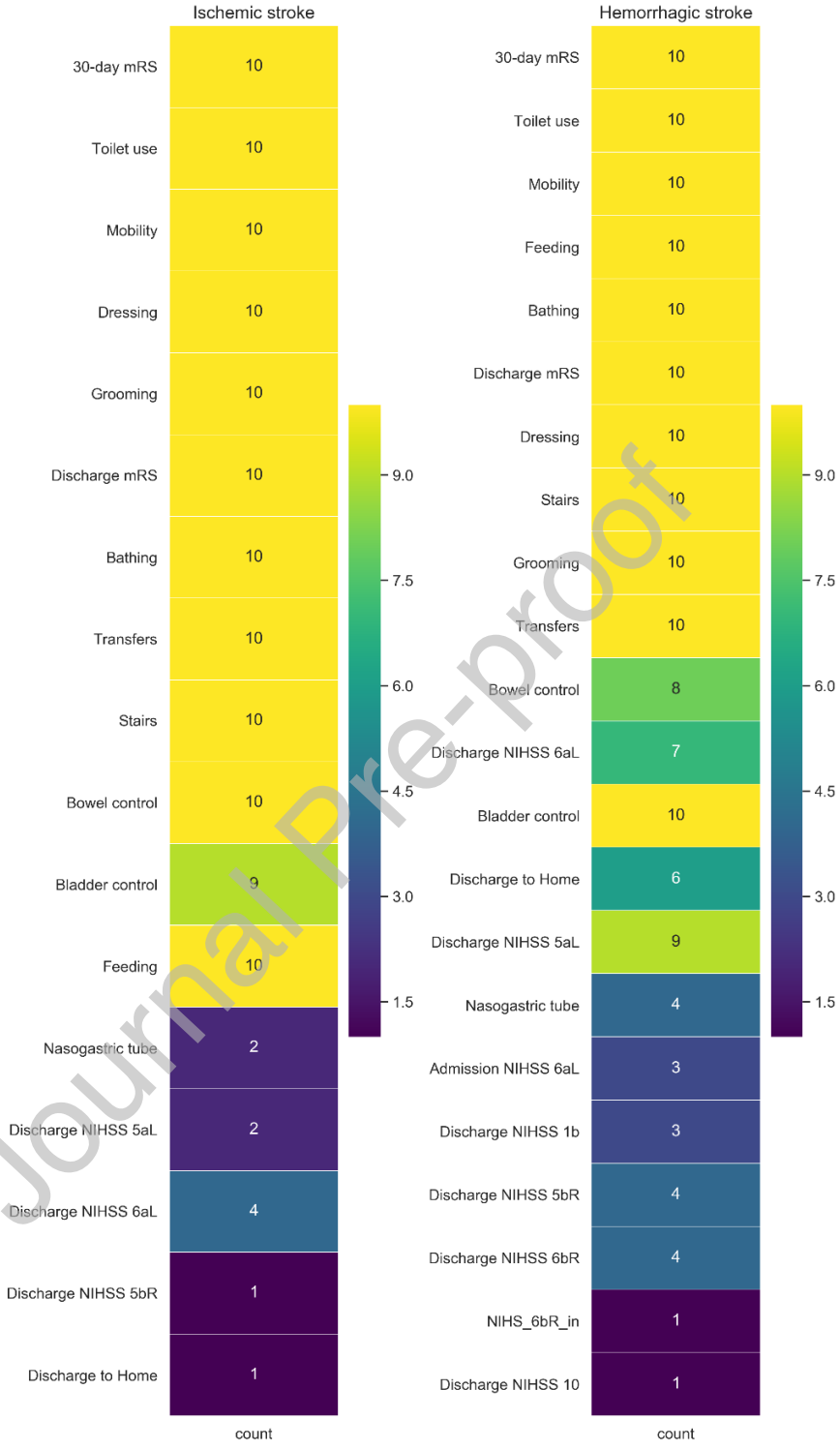


Figure 4. Selected features by extra-trees algorithm. The features are sorted by its importance. The numbers indicate how many times the feature passes the minimum threshold during cross-validation.

3. 3 Evaluation Results

Four supervised machine learning classifiers, SVM, random forest, ANN, and HANN, were evaluated on testing hold-out datasets repeated 10 times. The results of precision, recall and f1-score are presented in table 2. Using preadmission and inpatient data as feature inputs, the best performance was obtained in the ischemic stroke dataset with SVM (f1-score: 87.8%±0.2) and in the hemorrhagic stroke dataset with HANN (f1-score: 88.1%±0.7). In general, adding follow-up data which is the whole feature set, improves the performance of classifiers. The SVM achieved 92.9±0.1 f1-score in the ischemic stroke dataset and the random forest achieved 91.4±0.6 f1-score in the hemorrhagic stroke dataset. By using fewer selected features, the performance of all classifiers was not decreased but even slightly improved. The ANN classifier showed the worst performance in most cases, compared to other classifiers. Figure 5 shows the ROC and its area under the curve (AUC) of each classifier tested by hold-out datasets. For the ischemic dataset, the hybrid neural network techniques have better AUC results. The HANN has the highest AUC with whole feature set and selected feature set (0.974±0.000 and 0.971±0.001). SVM performs better than other classifiers in the hemorrhagic dataset (0.970±0.003 AUC with whole feature set and 0.973±0.002 AUC with selected feature set). Four classifiers show similar AUC results on the training datasets during 10-fold cross-validation (appendix figure B), which indicates that our classifiers are neither over-fitting nor under-fitting.

Table 2. The 10-time holdout testing result of machine learning models on 90-day stroke outcome prediction.

		Ischemic stroke (Mean ± SD) %			Hemorrhagic stroke (Mean ± SD) %		
		Precision	Recall	f1-score	Precision	Recall	f1-score
Clinical feature set (n=203)	ANN	86.5±0.2	89.2±0.3	87.6±0.6	87.9±0.9	88.3±1.0	88.0±0.9
	RF	86.6±0.2	89.8±0.2	87.8±0.2	87.7±0.7	88.4±0.8	87.9±0.8
	HANN	86.2±0.2	89.4±0.2	87.7±0.2	87.9±0.7	88.2±0.7	88.1±0.7
	SVM	86.7±0.2	89.4±0.2	87.8±0.2	87.0±0.8	87.4±0.8	87.1±0.8
Whole feature set (n= 206)	ANN	89.2±0.6	91.9±0.2	90.3±0.5	88.6±1.0	89.1±0.9	88.8±0.9
	RF	94.1±0.1	93.9±0.1	92.4±0.1	91.2±0.6	92.0±0.6	91.4±0.6
	HANN	91.5±0.5	93.9±0.3	92.6±0.4	89.0±0.7	89.5±0.6	89.2±0.7
	SVM	91.9±0.2	94.2±0.1	92.9±0.1	90.2±0.7	90.9±0.6	90.2±0.8
Selected feature set (n=17/22)	ANN	90.0±0.7	92.8±0.4	91.1±0.6	88.1±1.2	88.5±1.2	88.2±1.2
	RF	91.3±0.3	93.4±0.3	92.2±0.3	91.6±0.4	92.4±0.5	91.7±0.4
	HANN	91.8±0.3	94.1±0.1	92.8±0.2	90.1±0.4	90.7±0.6	90.1±0.4
	SVM	91.9±0.2	94.2±0.1	92.9±0.1	91.6±0.3	92.4±0.3	91.7±0.3

SVM: support vector machine, RF: random forest, ANN: artificial neural network, HANN: hybrid artificial neural network.
clinical feature set: preadmission and inpatient data, whole feature set: preadmission, inpatient and follow-up data, selected feature set: feature selection on whole feature set

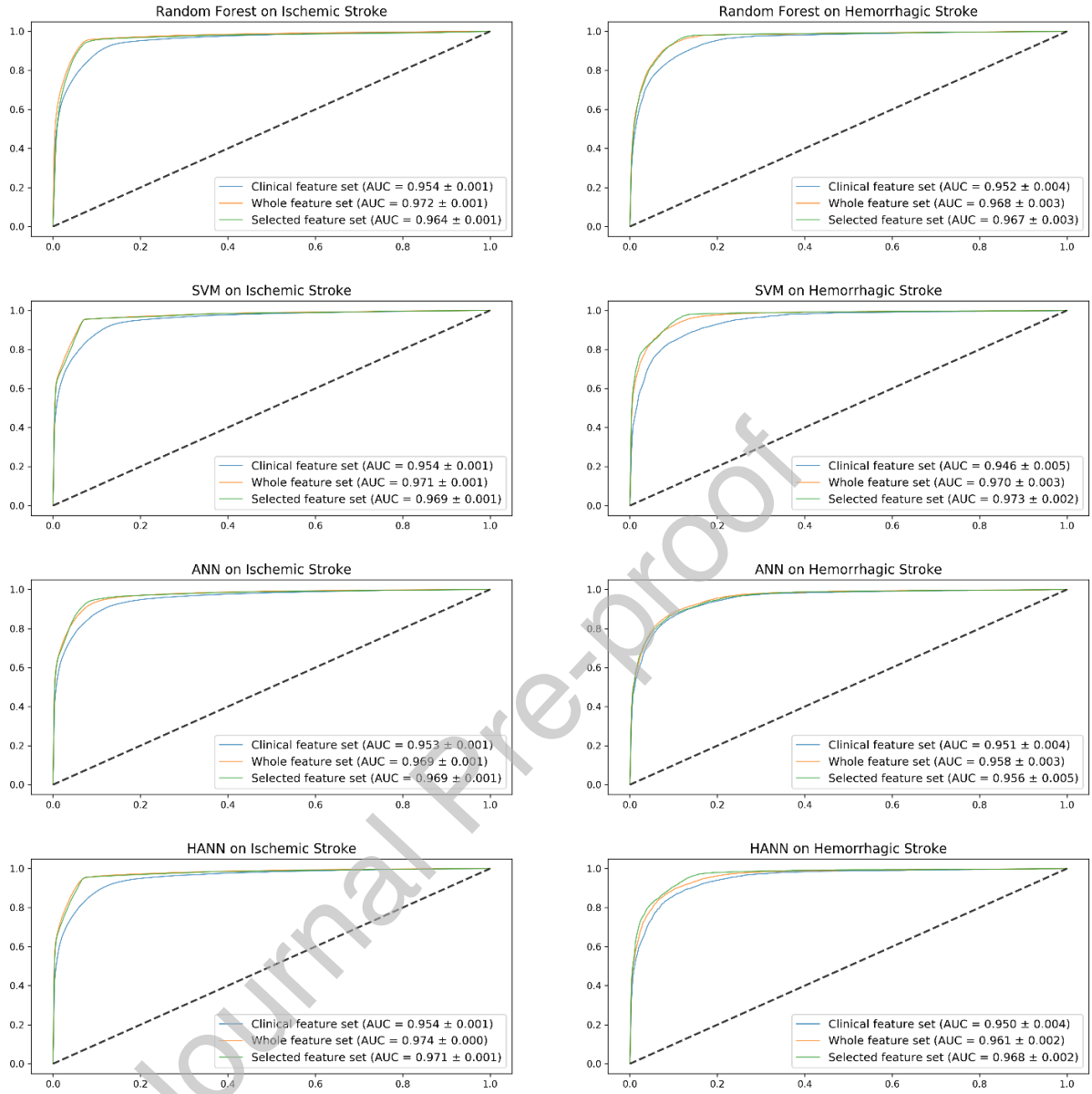


Figure 5. The receiver operating curve of 90-day stroke outcome prediction models on 10-times repeat hold-out testing data.

3.4 Misclassification Analysis

The evaluation results show that there is no significant performance difference between predictive models employed in this study. Therefore, we further analyzed those failed prediction cases with selected feature sets to see their distribution and characteristics. The Venn diagram shown in Figure 6 illustrates their distributions and characteristics. In all failed prediction cases, 32.1% cases (259/808) in the hemorrhagic stroke dataset and 55.2% cases (2,028/3,674) in the

ischemic stroke dataset were failed predicted by all four machine learning classifiers. The ratios were much higher than other intersections between two and three classifiers, in both stroke type. We compare the selected features of cases that misclassified by all classifiers (incorrect group) with other cases that correctly classified by all classifiers (correct group), the result shown in table 3 and table 4 by stroke type respectively. In both stroke types, there was no significant difference by gender, discharge to home, nasogastric tube insertion and NIHSS assessments. The average of 30-day mRS and discharge mRS of incorrect group are significantly higher than correct group in both stroke type. Except in bladder and bowel control, the differences of BI assessments in both groups also revealed that patients in incorrect group have more severe stroke which effects patient's daily activities.

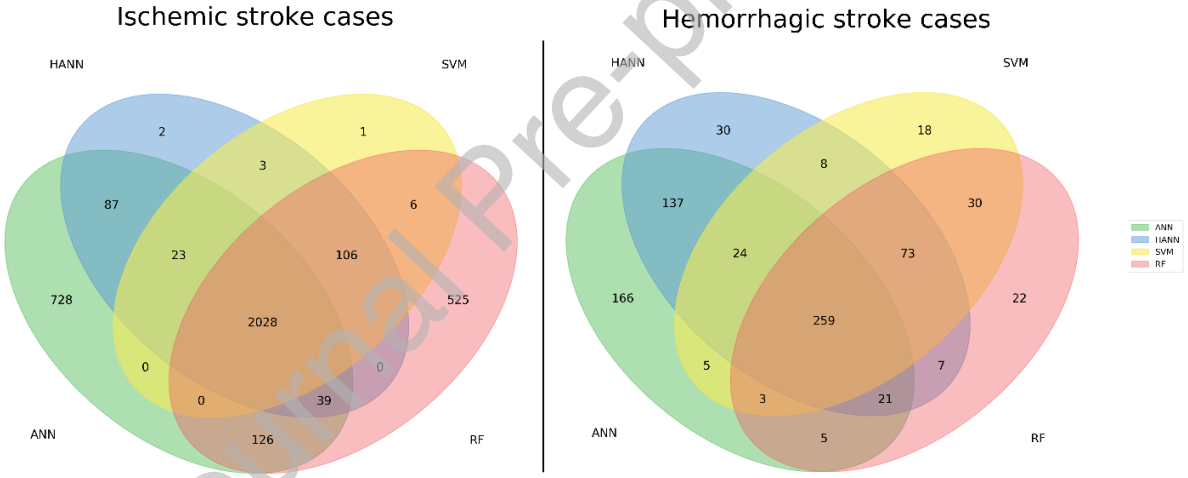


Figure 6. Venn diagram of failed prediction cases from four machine learning models.

Table 3. A comparison of misclassifications and correct classifications with demographics and selected features in ischemic stroke.

		Ischemic stroke			
		Misclassification (n=2,028)	**Correct classification (n=32,115)**	**P value**	**SMD**
Onset age (Mean ± SD)		68.91 ±11.78	67.67 ± 12.56	< 0.001	0.101
Gender (%)				0.799	0.006
Female		764 (37.7)	12,198 (38.0)		
Male		1,264 (62.3)	19,917 (62.0)		
90-day mRS (Mean ± SD)		2.25 ± 0.96	1.82 ± 1.63	< 0.001	0.321
30-day mRS (Mean ± SD)		2.75 ± 0.81	1.95 ± 1.57	< 0.001	0.634
Discharge mRS (Mean ± SD)		2.85 ± 0.85	2.14 ± 1.51	< 0.001	0.576
Transfers (%)				< 0.001	0.735
	0	23 (1.1)	2,852 (8.9)		
	5	162 (8.0)	2,757 (8.6)		
	10	980 (48.3)	5,904 (18.4)		
	15	863 (42.6)	20,602 (64.2)		
Toilet use (%)				< 0.001	0.84
	0	104 (5.1)	4,577 (14.3)		
	5	1,143 (56.4)	6,178 (19.2)		
	10	781 (38.5)	21,360 (66.5)		
Stairs (%)				< 0.001	0.898
	0	665 (32.8)	8,330 (25.9)		
	5	1,056 (52.1)	7,020 (21.9)		
	10	307 (15.1)	16,765 (52.2)		
Mobility (%)				< 0.001	0.65
	0	78 (3.8)	4,456 (13.9)		
	5	162 (8.0)	1,793 (5.6)		
	10	807 (39.8)	5,002 (15.6)		
	15	981 (48.4)	29,864 (65.0)		
Grooming (%)				< 0.001	0.741
	0	1,428 (70.4)	11,471 (35.7)		
	5	600 (29.6)	20,644 (64.3)		
Feeding (%)				< 0.001	0.472
	0	77 (3.8)	3,875 (12.1)		
	5	612 (30.2)	4,542(14.1)		
	10	1,339 (66.0)	23,698 (73.8)		
Dressing (%)				< 0.001	0.754
	0	75 (3.7)	3,941 (12.3)		
	5	1,059 (52.2)	6,221 (19.4)		
	10	894 (44.1)	21,953 (68.4)		

Table 3 (cont.) A comparison of misclassifications and correct classifications with demographics and selected features in ischemic stroke.

		Ischemic stroke			
		Misclassification (n=2,028)	**Correct classification (n=32,115)**	**P value**	**SMD**
Bathing (%)				< 0.001	0.124
	0	599 (29.5)	7,727 (24.1)		
	5	1,429 (70.5)	24,388 (75.9)		
Bladder control (%)				< 0.001	0.339
	0	44 (2.2)	3,275 (10.2)		
	5	105 (5.2)	1,397 (4.3)		
	10	1,879 (92.7)	27,443 (85.5)		
Bowel control (%)				< 0.001	0.358
	0	27 (1.3)	2,943 (9.2)		
	5	85 (4.2)	1,402 (4.4)		
	10	1,916 (94.5)	27,770 (86.5)		
Discharge to Home (%)				0.995	0.001
	No	184 (9.1)	2,904 (9.0)		
	Yes	1,844 (90.9)	29,211 (91.0)		
Nasogastric tube (%)				0.058	0.045
	No	1,757 (86.6)	27,319 (85.1		
	Yes	271 (13.4)	4,796 (14.9)		
Discharge NIHSS 5aL (Mean ± SD)		0.41 ± 0.73	0.46 ± 0.97	0.019	0.060
Discharge NIHSS 6aL (Mean ± SD)		0.4 ± 0.63	0.47 ± 0.93	0.002	0.084
Discharge NIHSS 5bR (Mean ± SD)		0.4 ± 0.72	0.45 ± 0.96	0.01	0.065

Table 4. A comparison of misclassifications and correct classifications with demographics and selected features in hemorrhage stroke.

	Hemorrhage stroke			
	Misclassification (n=259)	**Correct classification (n=3,806)**	**P value**	**SMD**
Onset age (Mean ± SD)	59.95 ± 12.44	61.4 ± 13.79	0.1	0.11
Gender (%)			0.64	0.034
Female	162 (37.5)	1,489 (39.1)		
Male	162 (62.5)	2,317 (60.9)		
90-day mRS (Mean ± SD)	2.14 (0.85)	2.29 (1.76)	0.17	0.11
30-day mRS (Mean ± SD)	2.98 (0.72)	2.53 (1.72)	< 0.001	0.34
Discharge mRS (Mean ± SD)	3.24 (0.91)	2.76 (1.59)	< 0.001	0.37
Transfers (%)			< 0.001	0.72
0	13 (5.0)	666 (17.5)		
5	41 (15.8)	512 (13.5)		
10	120 (46.3)	709 (18.6)		
15	85 (32.8)	1,919 (50.4)		
Toilet use (%)			< 0.001	0.89
0	43 (16.6)	1,052 (27.6)		
5	159 (61.4)	823 (21.6)		
10	57 (22.0)	1,931 (50.7)		
Stairs (%)			< 0.001	0.71
0	134 (51.7)	1,599 (42.0)		
5	95 (36.7)	706 (18.5)		
10	30 (11.6)	1,501 (39.4)		
Mobility (%)			< 0.001	0.66
0	34 (13.1)	1,028 (27.0)		
5	40 (15.4)	328 (8.6)		
10	92 (35.5)	500 (13.1)		
15	93 (35.9)	1,950 (51.2)		
Grooming (%)			< 0.001	0.78
0	225 (86.9)	2,048 (53.8)		
5	34 (13.1)	1,757 (46.2)		
Feeding (%)			< 0.001	0.67
0	20 (7.7)	840 (22.1)		
5	122 (47.1)	733 (19.3)		
10	117 (45.2)	2,233 (58.7)		
Dressing (%)			< 0.001	0.84
0	25 (9.7)	879 (23.1)		
5	160 (61.8)	907 (23.8)		
10	74 (28.6)	2,020 (53.1)		

Table 4 (cont.) A comparison of misclassifications and correct classifications with demographics and selected features in hemorrhage stroke.

		Hemorrhage stroke			
		Misclassification (n=259)	**Correct classification (n=3,806)**	**P value**	**SMD**
Bathing				0.08	0.116
	0	114 (44.0)	1,459 (38.3)		
	5	145 (56.0)	2,347 (61.7)		
Bladder control				< 0.001	0.31
	0	26 (10.0)	769 (20.2)		
	5	28 (10.8)	250 (6.6)		
	10	205 (79.2)	2,787 (73.2)		
Bowel control				< 0.001	0.379
	0	15 (5.8)	671 (17.6)		
	5	24 (9.3)	253 (6.6)		
	10	220 (84.9)	2,882 (75.7)		
Discharge to Home				0.58	0.042
	No	49 (18.9)	782 (20.6)		
	Yes	210 (81.1)	3,023 (79.4)		
Nasogastric tube				0.407	0.06
	No	176 (67.9)	2,482 (65.2)		
	Yes	83 (32.0)	1,324 (34.8)		
Discharge NIHSS 1b (Mean ± SD)		0.20 ± 0.52	0.35 ± 0.71	0.001	0.237
Discharge NIHSS 5aL (Mean ± SD)		0.64 ± 0.99	0.76 ± 1.26	0.136	0.105
Discharge NIHSS 6bR (Mean ± SD)		0.58 (0.38)	0.73 (1.2)	0.055	0.141
Discharge NIHSS 6aL (Mean ± SD)		0.64 (0.96)	0.78 (1.24)	0.079	0.125
Discharge NIHSS 5bR (Mean ± SD)		0.6 (0.91)	0.7 (1.2)	0.197	0.092
Discharge NIHSS 10 (Mean ± SD)		0.44 (0.54)	0.50 (0.70)	0.148	0.103
Admission NIHSS 6aR (Mean ± SD)		1.01 (1.22)	0.97 (1.33)	0.629	0.032
Admission NIHSS 6aL (Mean ± SD)		1.05 ± 1.26	1.02 ± 1.34	0.72	0.024

4. Discussion

Improving the outcome of stroke is a global priority. Outcome prediction plays an important role in evidence-based decision-making and guides clinicians on how to best treat stroke patients. Medical registries have been used for years as sources of clinical data to support

evidence-based medicine. This study developed models of excellent stroke outcome prediction by using the Taiwan Stroke Registry (TSR).

It is difficult to accurately predict functional outcomes after stroke [31] even for experienced clinicians. Patients with less initial motor impairment [32] and less corticomotor system defects [33] might have better motor outcomes. However, these correlations are not always correct when predicting the prognoses. There are numerous factors, including clinical features and treatments, that can influence the final stroke outcome. Consequently, these complicated interactions make conventional modelling very challenging to predict outcomes. Machine learning models are relatively independent of the underlying interactions and are able to simulate the result of a complex system. Many studies have proposed prognostic models for stroke outcome prediction using machine learning or statistic approach. Asadi et al. developed dichotomized mRS models of acute ischemic stroke and presented 0.6 AUC of ANN and 70% accuracy of SVM on a small clinical dataset (107 cases) [34]. In imaging-based machine learning for predicting stroke outcomes, Bentley et al. built an SVM model to identify acute ischemic stroke patients at risk for symptomatic intracranial hemorrhage using 116 acute ischemic stroke patients' CT brain images. The AUC of their prognostic model achieved 0.744 [35]. Muscari et al. constructed a simple prognostic scale called Bologna Outcome Algorithm for Stroke (BOAS) to predicate dependency or death after ischemic stroke based on 221 ischemic stroke patients. In the test group, the accuracy was 79.0% and the AUC was 0.839 [36]. Heo et al. developed machine learning-based models with a prospective cohort of 2,923 patients with acute ischemic stroke. The AUC was 0.888 for ANN model, 0.810 for random forest model, 0.836 for SVM model, and 0.842 for logistic regression model [37].

Compared to previous studies, our machine learning classifiers present relatively high performance (approximate 0.95 AUC) in both ischemic and hemorrhagic stroke. There are several factors considered to be causally associated with our higher performance. The first factor is the data quantity which is important for the machine learning algorithms. The supervised machine learning requires a large amount of labeled data to optimize its model. This study applied machine learning algorithms on 35,798 ischemic stroke cases and 4,495 hemorrhagic cases, which provides a sample size is significantly larger than the other studies mentioned above. The second factor is the data quality. The inherent need for large training datasets may affect the accuracy of the machine learning algorithms in studies. Poor data quality including outliers,

wrong labeling, or conflicting data can mislead the machine learning process. To avoid this, we used a series of data cleaning and validation processes ensure the quality of stroke datasets obtained from the national wide stroke registry repository for model training and testing. For data validation, we employed clinical-logic validation and LOWESS regulation by identifying the inconsistency of mRS, Barthel index, and NIHSS. This method can also be applied to other correlated features in the data bank based on the understanding of clinical meanings. The performance results also reveal that follow-up data is very useful for out-come prediction. Taking SVM as an example, it improves the AUC from 0.954 to 0.971 in ischemic stroke and from 0.946 to 0.970 in hemorrhage stroke. The follow-up data includes the 30-day mRS, which has the highest importance in both stroke type dataset. Previous studies have also reported similar finding that the 30-day mRS is highly correlated with 90-day outcome [38, 39].

The AUC results calculated from 10-times repeat hold-out testing data found no significant performance differences among four machine learning classifiers. Therefore, the most important consideration may be the algorithm's characteristics when choosing which machine learning model should be applied in medicine. For example, the random forest is able to report the feature's importance, the neural network may need to be designed with an appropriate network structure based on the data characteristics using appropriate layers to process the data with different characteristics. It greatly improves the ability of neural network to extract patterns in data. In a real-world medical scenario, electronic medical records accumulate in hospital information system and it is computationally infeasible to train over the entire dataset all the time. Compared to traditional machine learning approach in a batch learning setting, online machine learning is a fundamentally different approach that update models from data streams sequentially [40]. The online machine learning can be directly applied to backpropagation neural networks [41], therefore the neural network can be a better choice if we expect our prediction model to be updated. According to the misclassification analysis result, the accuracy of predicting the more severe stroke patent's outcome could be a discriminatory measure, if the major issue is model's performance on outcome prediction.

In previous studies, stroke outcomes were influenced by stroke severity [7, 8], age [7, 8, 42, 43], sex [42-44] and comorbidities [44]. We performed feature selection in this study and 17 features were selected for ischemic stroke and 22 features were selected for hemorrhagic stroke from a total 206 features. The results showed that we identified a much smaller set of input

features that can still achieve good predictive performance. The benefit of the feature selection makes the machine learning classifier more practical because if fewer variables were required then it will reduce data entry time and expands the data generation. Among these selected features, behavior assessments (30-day mRS and the Barthel index) and neurological assessments (NIHSS) are selected as the most important features for prediction models. We calculated the correlation between 90-day mRS outcome and the total score of assessments using Spearman correlation. It shows strong positive correlation (0.92) between 30-day mRS and 90-day mRS, strong negative correlation (-0.79) between Barthel index and 90-day mRS. The correlation between NIHSS and 90-day mRS is moderate correlation (0.66). The correlations reveal that behavior and neurological assessments are informative features for the ML algorithms in 90-day mRS outcome prediction task. Some managements during admission were also selected and the odds ratio was calculated for these features. The odds ratios (OR) of nasogastric (NG) tube insertion in ischemic and hemorrhagic stroke patients are 11.5 (10.8-12.4) and 7.2 (6.3-8.3), respectively, and the results may imply that NG tubes are required in more severe patients. The OR of discharge to home in ischemic and hemorrhagic stroke patients are 0.056 (0.051-0.062) and 0.076 (0.063-0.093), respectively, and that may suggest that less severe patients can care themselves at home without transferring to other facilities. These selected features are adapted from TSR recording and relatively easy to obtain compared to image data. Consequently, this model is more practical for clinical use by physicians.

There are still some limitations about the machine learning models. For example, these classifiers although accurate in predicting the functional outcome of stroke behave as a black box [45, 46]. Clinicians cannot explain the relationship between the input features and the outcome to patients. Also, this study only compared machine learning algorithms on two types of stroke in a general scenario. Evaluating various machine learning classifiers by considering other moderator effects (i.e. age, gender) and specific clinical pathway (i.e. in an emergency room) will be explored further. All classifiers were not applied hyperparameter optimization.

5. Conclusion

In the present study, we assessed various supervised machine learning classifiers of their capability in 90-day mRS outcome prediction based on a national stroke registry. The results revealed that applying machine learning algorithms on a large dataset for important feature

selection and classifier training can be a powerful tool for stroke outcome prediction. The follow-up data is very useful for outcome prediction. In the case of the classifiers with similar performance, we can choose the predictive model based on the real-world requirements and its performance on severe stroke patients.

Declaration of Competing Interest: The authors have no competing interests to declare.

Acknowledgements: The authors thank the NIH library editing service. This study is supported in part by Taiwan Ministry of Health and Welfare Clinical Trial Center (MOHW108-TDU-B-212-133004), China Medical University Hospital, Academia Sinica Stroke Biosignature Project (BM10701010021), MOST Clinical Trial Consortium for Stroke (MOST 107-2321-B-039 -004-), Tseng-Lien Lin Foundation, Taichung, Taiwan, and Katsuzo and Kiyo Aoshima Memorial Funds, Japan. Authors wish to express our sincere gratitude to those TSR investigators listed in appendix without whom this study would not be possible.

Contributors: CHL conceived the idea of the study, implemented the machine learning approaches and drafted the manuscript. KCH and KRJ performed the statistic analyze and interpreted the result. CYH provided practical suggestion to this study. CHT, YS, LML, WLC, PLC, CLL and CYH processed and provided dataset. YCF provided key support, coordinated cooperative organizations and input practical concerns to the study. All authors contributed to the review of the manuscript and approved the final version.

Competing interests Statement: the authors have no competing interests to declare.

Reference

[1] Krishnamurthi R, Feigin V, Forouzanfar M, Mensah G, Connor M, Bennett D, et al. Global Burden of Diseases, Injuries, Risk Factors Study 2010 (GBD 2010); GBD Stroke Experts Group. Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet Glob Health. 2013;1:e259-e81.

[2] Thrift AG, Cadilhac DA, Thayabaranathan T, Howard G, Howard VJ, Rothwell PM, et al. Global stroke statistics. International Journal of Stroke. 2014;9:6-18.

[3] Barker-Collo S, Feigin V, Parag V, Lawes C, Senior H. Auckland stroke outcomes study: part 2: cognition and functional outcomes 5 years poststroke. Neurology. 2010;75:1608-16.
[4] Bates BE, Xie D, Kwong PL, Kurichi JE, Ripley DC, Davenport C, et al. Development and validation of prognostic indices for recovery of physical functioning following stroke: part 2. PM&R. 2015;7:699-710.
[5] Bates BE, Xie D, Kwong PL, Kurichi JE, Ripley DC, Davenport C, et al. Development and validation of prognostic indices for recovery of physical functioning following stroke: part 1. PM&R. 2015;7:685-98.
[6] Stinear C. Prediction of recovery of motor function after stroke. The Lancet Neurology. 2010;9:1228-32.
[7] Meyer MJ, Pereira S, McClure A, Teasell R, Thind A, Koval J, et al. A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation. Disability and rehabilitation. 2015;37:1316-23.
[8] Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW. Early prediction of outcome of activities of daily living after stroke: a systematic review. Stroke. 2011;42:1482-8.
[9] Saposnik G, Ntaios G, Michel P, Team iR. An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL scoreAuthor Response. Neurology. 2012;79:2293-4.
[10] Strbian D, Meretoja A, Ahlhelm F, Pitkäniemi J, Lyrer P, Kaste M, et al. Predicting outcome of IV thrombolysis–treated ischemic stroke patients The DRAGON score. Neurology. 2012;78:427-32.
[11] Flint A, Cullen S, Faigeles B, Rao V. Predicting long-term outcome after endovascular stroke treatment: the totaled health risks in vascular events score. American Journal of Neuroradiology. 2010;31:1192-6.
[12] Deo RC. Machine learning in medicine. Circulation. 2015;132:1920-30.
[13] Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine. 2016;375:1216.

[14] Hsieh F-I, Lien L-M, Chen S-T, Bai C-H, Sun M-C, Tseng H-P, et al. Get with the guidelines-stroke performance indicators: surveillance of stroke care in the Taiwan stroke registry: get with the guidelines-stroke in Taiwan. Circulation. 2010;122:1116-23.
[15] Mohanty C, Ray S, Singhal A. Relationship between Barthel Index (BI) and the Modified Rankin Scale (mRS) Score in Assessing Functional Outcome in Acute Ischemic Stroke. Journal of Marine Medical Society. 2016;18:144.
[16] Cioncoloni D, Piu P, Tassi R, Acampa M, Guideri F, Taddei S, et al. Relationship between the modified Rankin Scale and the Barthel Index in the process of functional recovery after stroke. NeuroRehabilitation. 2012;30:315-22.
[17] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association. 1979;74:829-36.
[18] Khatri P, Abruzzo T, Yeatts S, Nichols C, Broderick J, Tomsick T. Good clinical outcome after ischemic stroke with successful revascularization is time-dependent. Neurology. 2009;73:1066-72.
[19] Castellanos M, Leira R, Tejada J, Gil-Peralta A, Davalos A, Castillo J. Predictors of good outcome in medium to large spontaneous supratentorial intracerebral haemorrhages. Journal of Neurology, Neurosurgery & Psychiatry. 2005;76:691-5.
[20] Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. Stroke. 1999;30:1538-41.
[21] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods: Cambridge university press; 2000.
[22] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2:18-22.
[23] Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment. 1998;32:2627-36.

[24] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12:2825-30.
[25] Keras. Available at <https://keras.io>. [accessed 24 Sep 2018].
[26] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63:3-42.
[27] Kouwaye B. Regression Trees and Random forest based feature selection for malaria risk exposure prediction. arXiv preprint arXiv:160607578. 2016.
[28] Wilson JL, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW, et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. Stroke. 2002;33:2243-6.
[29] Shah S, Vanclay F, Cooper B. Improving the sensitivity of the Barthel Index for stroke rehabilitation. Journal of clinical epidemiology. 1989;42:703-9.
[30] Hage V. The NIH stroke scale: a window into neurological status. NurseCom Nursing Spectrum (Greater Chicago). 2011;24:44-9.
[31] Nijland RH, Van Wegen EE, Harmeling-van der Wel BC, Kwakkel G, Investigators EPoFOAS. Accuracy of physical therapists' early predictions of upper-limb function in hospital stroke units: the EPOS Study. Physical therapy. 2013;93:460-9.
[32] Coupar F, Pollock A, Rowe P, Weir C, Langhorne P. Predictors of upper limb recovery after stroke: a systematic review and meta-analysis. Clinical rehabilitation. 2012;26:291-313.
[33] Kim B, Winstein C. Can neurological biomarkers of brain impairment be used to predict poststroke motor recovery? A systematic review. Neurorehabilitation and neural repair. 2017;31:3-24.
[34] Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PloS one. 2014;9:e88225.

[35] Bentley P, Ganesalingam J, Jones ALC, Mahady K, Epton S, Rinne P, et al. Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage: Clinical. 2014;4:635-40.
[36] Muscari A, Puddu G, Santoro N, Zoli M. A simple scoring system for outcome prediction of ischemic stroke. Acta neurologica Scandinavica. 2011;124:334-42.
[37] Heo J, Yoon J, Park HJ, Kim YD, Nam HS, Heo JH. Machine Learning-Based Model Can Predict Stroke Outcome. Stroke. 2018;49:A194-A.
[38] Ovbiagele B, Lyden PD, Saver JL. Disability status at 1 month is a reliable proxy for final ischemic stroke outcome. Neurology. 2010;75:688-92.
[39] Rost NS, Bottle A, Lee JM, Randall M, Middleton S, Shaw L, et al. Stroke severity is a crucial predictor of outcome: an international prospective validation study. Journal of the American Heart Association. 2016;5:e002433.
[40] Shalev-Shwartz S, Singer Y. Online learning: Theory, algorithms, and applications. 2007.
[41] Sahoo D, Pham Q, Lu J, Hoi SC. Online deep learning: Learning deep neural networks on the fly. arXiv preprint arXiv:171103705. 2017.
[42] Kelly-Hayes M, Beiser A, Kase CS, Scaramucci A, D'Agostino RB, Wolf PA. The influence of gender and age on disability following ischemic stroke: the Framingham study. Journal of Stroke and Cerebrovascular Diseases. 2003;12:119-26.
[43] Nichols-Larsen DS, Clark P, Zeringue A, Greenspan A, Blanton S. Factors influencing stroke survivors' quality of life during subacute recovery. Stroke. 2005;36:1480-4.
[44] Reeves MJ, Bushnell CD, Howard G, Gargano JW, Duncan PW, Lynch G, et al. Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. The Lancet Neurology. 2008;7:915-26.
[45] Alpaydin E. Introduction to machine learning: MIT press; 2009.
[46] Warwick K. Artificial intelligence: the basics: Routledge; 2013.

Appendix A: List of Taiwan Stroke Registry Investigators

China Medical University Hospital: Yuh-Cherng Guo (Principal Investigator), Chon-Haw Tsai, Wei-Shih Huang, Chung-Ta Lu, Tzung-Chang Tsai, Chun-Hung Tseng, Kang-Hsu Lin, Woei-Cherng Shyn, Yu-Wan Yang, Yen-Liang Liu, Der-Yang Cho, Chun-Chung Chen, Chung-Hsiang Liu

National Taiwan University Hospital: Jiann-Shing Jeng (Principal Investigator), Sung-Chun Tang, Li-Kai Tsai, Shin-Joe Yeh

E-Da Hospital / I-Shou University: Shih-Pin Hsu (Principal Investigator), Han-Jung Chen, Cheng-Sen Chang, Hung-Chang Kuo, Lian-Hui Lee, Huan-Wen Tsui, Jung-Chi Tsou, Yan-Tang Wang, Yi-Cheng Tai,Kun-Chang Tsai, Yen-Wen Chen, Kan Lu, Po-Chao Liliang, Yu-Tun Tsai, Cheng-Loong Liang, Kuo-Wei Wang, Hao-Kuang Wang, Jui-Sheng Chen, Po-Yuan Chen, Cien-Leong Chye, Wei-Jie Tzeng, Pei-Hua Wu

National Cheng Kung University Hospital: Chih-Hung Chen (Principal Investigator), Pi-Shan Sung, Han-Chieh Hsieh, Hui-Chen Su

Shin Kong WHS Memorial Hospital: Hou-Chang Chiu (Principal Investigator), Li-Ming Lien, Wei-Hung Chen, Chyi-Huey Bai, Tzu-Hsuan Huang, Chi-Ieong Lau,Ya-Ying Wu, Hsu-Ling Yeh, Anna Chang

Kaohsiung Veterans General Hospital: Ching-Huang Lin (Principal Investigator), Cheng-Chang Yen

Kaohsiung Medical University Chung-Ho Memorial Hospital: Ruey-Tay Lin (Principal Investigator), Chun-Hung Chen, Gim-Thean Khor, A-Ching Chao, Hsiu-Fen Lin, Poyin Huang

Chi Mei Medical Center: Huey-Juan Lin (Principal Investigator), Der-Shin Ke, Chia-Yu Chang, Poh-Shiow Yeh, Kao-Chang Lin, Tain-Junn Cheng, Chih-Ho Chou, Chun-Ming Yang, Hsiu-Chu Shen

Chung Shan Medical University Hospital: An-Chih Chen (Principal Investigator), Shih-Jei Tsai, Tsong-Ming Lu, Sheng-Ling Kung, Mei-Ju Lee, Hsi-Hsien Chou

Show Chwan Memorial Hospital: Hsin-Yi Chi (Principal Investigator), Chou-Hsiung Pan, Po-Chi Chan, Min-Hsien Hsu, Wei-Lun Chang,Ya-Ying Wu , Zhi-Zang Huang , Hai-Ming Shoung,Yi-Chen Lo, Fu-Hwa Wang

Cheng Hsin General Hospital: Ta-Chang Lai (Principal Investigator), Jiu-Haw Yin, Chung-JenWang, Kai-ChenWang, Li-Mei Chen, Jong-Chyou Denq

En Chu Kong Hospital: Yu Sun (Principal Investigator), Chien-Jung Lu, Cheng-Huai Lin, Chieh-Cheng Huang, Chang-Hsiu Liu, Hoi-Fong Chan

Far Eastern Memorial Hospital: Siu-Pak Lee (Principal Investigator)

Kuang Tien General Hospital: Ming-Hui Sun (Principal Investigator), Li-Ying Ke

Taichung Veterans General Hospital: Po-Lin Chen (Principal Investigator), Yu-Shan Lee

Ditmanson Medical Foundation Chia-Yi Christian Hospital: Sheng-Feng Sung(Principal Investigator), Cheung-Ter Ong, Chi-Shun Wu, Yung-Chu Hsu, Yu-Hsiang Su, Ling-Chien Hung

Tri-Service General Hospital: Jiunn-Tay Lee (Principal Investigator), Jiann-Chyun Lin, Yaw-Don Hsu, Jong-Chyou Denq, Giia-Sheun Peng, Chang-Hung Hsu, Chun-Chieh Lin, Che-Hung Yen, Chun-An Cheng, Yueh-Feng Sung, Yuan-Liang Chen, Ming-Tung Lien, Chung-Hsing Chou, Chia-Chen Liu, Fu-Chi Yang, Yi-ChungWu, An-Chen Tso, Yu- Hua Lai, Chun-I Chiang, Chia-Kuang Tsai, Meng-Ta Liu, Ying-Che Lin, Yu-Chuan Hsu

Cathay General Hospital: Tsuey-Ru Chiang (Principal Investigator), Mei-Ching Lee, Pai-Hao Huang, Sian-King Lie, Pin-Wen Liao, Jen-Tse Chen

Changhua Christian Hospital: Mu-Chien Sun (Principal Investigator), Tien-Pao Lai, Wei-Liang Chen, Yen-Chun Chen, Ta-Cheng Chen, Wen-Fu Wang, Kwo-Whei Lee, Chen-Shu Chang, Chien-Hsu Lai, Siao-Ya Shih, Chieh-Sen Chuang, Yen-Yu Chen, Chien-Min Chen

Taipei Tzuchi Hospital: Shinn-Kuang Lin (Principal Investigator, School of Medicine, Tzuchi University, Hualien, Taiwan), Yu-Chin Su, Cheng-Lun Hsiao, Fu-Yi Yang, Chih-Yang Liu, Han-Lin Chiang.

Min Sheng General Hospital: Chun-Yuan Chang (Principal Investigator), I-sheng Lin,Chung-Hsien Chien,Yang-Chuang Chang

Lin Shin Hospital: Ping-Kun Chen (Principal Investigator), Pai-Yi Chiu

National Taiwan University Hospital Yunlin Branch: Yu-Jen Hsiao (Principal Investigator), Chen-Wen Fang

Landseed Hospital: Yu-Wei Chen (Principal Investigator), Kuo-Ying Lee, Yun-Yu Lin, Chen-Hua Li, Hui-Fen Tsai, Chuan-Fa Hsieh, Chih-Dong Yang, Shiumn-Jen Liaw, How-Chin Liao

Cheng Ching General Hospital: Shoou-Jeng Yeh (Principal Investigator), Ling-Li Wu, Liang-Po Hsieh, Yong-Hui Lee, Chung-Wen Chen

China Medical University Beigang Hospital: Chih-Shan Hsu(Principal Investigator),Ye-Jian-Jhih, Hao-Yu Zhuang, Yan-Hong Pan, Shin-An Shih

Taipei Medical University -Wan Fang Hospital: Chin-I Chen (Principal Investigator), Jia-Ying Sung, , Hsing-Yu Weng, Hao-Wen Teng, Jing-Er Lee, Chih-Shan Huang, Shu-Ping Chao

Taipei Medical University Hospital: Rey-Yue Yuan (Principal Investigator), , Jau- Jiuan Sheu, Jia-Ming Yu, Chun-Sum Ho, Ting-Chun Lin

Kuang Tien General Hospital Dajia Division: Shih-Chieh Yu(Principal Investigator)

Changhua Christian Hospital Yunlin Branch: Jiunn-Rong Chen (Principal Investigator), Song-Yen Tsai

Chang Bing Show Chwan Memorial Hospital: Cheng-Yu Wei (Principal Investigator), Tzu-Hsuan Huang, Chao-Nan Yang, Chao-Hsien Hung, Ian Shih

Lotung Poh Ai Hospital: Hung-Pin Tseng (Principal Investigator), Chin-Hsiung Liu, Chun-Liang Lin, Hung-Chih Lin, Pi-Tzu Chen

Taipei Medical University - Shuang Ho Hospital: Chaur-Jong Hu (Principal Investigator), Nai-Fang Chi, Lung Chan

Taipei Veterans General Hospital & National Yang-Ming University School of Medicine: Chang-Ming Chern (Principal Investigator), Chun-Jen Lin, Shuu-Jiun Wang, Li-Chi Hsu, Wen-Jang Wong, I-Hui Lee, Der-Jen Yen, Ching-Piao Tsai, Shang-Yeong Kwan, Bing-Wen Soong, Shih-Pin Chen, Kwong-Kum Liao, Kung-Ping Lin, Chien Chen, Din-E Shan, Jong-Ling Fuh, Pei-Ning Wang, Yi-Chung Lee, Yu-Hsiang Yu, Hui-Chi Huang, Jui-Yao Tsai

Chi Mei Medical Center, Liouying: Ming-Hsiu Wu (Principal Investigator), Shi-Cheng Chen, Szu-Yi Chiang, Chiung-Yao Wang

Buddhist Dalin Tzu Chi General Hospital: Ming-Chin Hsu (Principal Investigator)

St. MARTIN DE PORRES HOSPITAL: Chien-Chung Chen (Principal Investigator), Po-Yen Yeh, Yu-Tai Tsai, Ko-Yi Wang
Sin-Lau Hospital, Tainan, the Presbyterian Church in Taiwan: Tsang-Shan Chen(Principal Investigator)
Cardinal Tien Hospital: Ping-Keung Yip (Principal Investigator), Vinchi Wang, Kaw-ChenWang, Chung-Fen Tsai, Chao-Ching Chen, Chih-Hao Chen, Yi-Chien Liu, Shao-Yuan Chen, Zi-Hao Zhao, Zhi-Peng Wei
Yumin Medical Corporation Yumin Hospital: Shey-Lin Wu(Principal Investigator)
Kaohsiung Municipal Hsiao-kang Hospital: Ching-Kuan Liu(Principal Investigator)
Wei Gong Memorial Hospital: Ryh-Huei Lin (Principal Investigator), Ching-Hua Chu
Taipei City Hospital Ren Ai Branch: Sui-Hing Yan (Principal Investigator), Yi-Chun Lin, Pei-Yun Chen, Sheng-Huang Hsiao
National Taiwan University Hospital Hsin-Chu Branch: Bak-Sau Yip (Principal Investigator), Pei-Chun Tsai,Ping-Chen Chou, Tsam-Ming Kuo, Yi-Chen Lee, Yi-Pin Chiu, Kun-Chang Tsai
Taichung Hospital Department of Health : Yi-Sheng Liao (Principal Investigator)
Tainan Municipal An-Nan Hospital-China Medical University: Ming-Jun Tsai (Principal Investigator), Hsin-Yi Kao

Appendix B: Details of Machine learning models

Support vector machine

The SVM is considered as one of the most robust and accurate method among the well-known machine learning algorithms and has been widely used in clinical outcome prediction. We used the linear kernel function with a default penalty parameter C = 0.1.

Random forest

Random forest uses an ensemble of classification trees and works well with a mixture of numerical and categorical features. There are three primary parameters of Random forest classifier. The *n_estimators* is the number of trees we want to build before taking the maximum voting of predictions. In this study, we set *n_estimators* as default value 100. Second, the number of features (*max_features*) is the maximum number of features that random forest allowed to try in an individual tree. We set it as the square root of number of input features. Finally, the Gini impurity: $1 - \sum_{i=1}^{j} p_i^2$, where j is the number of classes and $p_i$be the fraction of items labeled with class i in dataset, was selected for the function to measure the quality of a split.

Artificial neural network

We built a basic three-layer neural network that contains an input layer, one hidden layer of 2/3 the size of the input layer neurons with the rectification nonlinearity (ReLU) activation function, given by $f(x) = \max(0, x)$ and an output layer of two neurons with the softmax function:$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{c} y_j}$, where j is the number of output classes. To train the neural network, the loss function and the network optimizer were set as binary cross entropy and stochastic gradient descent (SGD) with the learning rate $5e^{-3}$ respectively. The neural network training was done for 150 epochs with batch size of 56.

Hybrid artificial neural network

The convolutional neural network (CNN) is a powerful technology and has achieved remarkable results in recent years. The core building block of a CNN is a set of learnable convolving filters in a convolutional layer that can calculate the inner product between the filter and input to detect some specific type of feature over a single spatial or temporal input. Inspired by the behavior of the convolutional layer, this study designed a hybrid neural network model that combined a dot product layer and a multi-layer perceptron (MLP) neural network to identify the pattern of various types of clinical data. We separated the clinical data input by whether it is

temporal data or not. For example, the admission and discharge NIHSS assessments as well as the discharge and 30-day mRS follow up are considered as temporal related data input. The rest of clinical data, such as demographic, laboratory data, and medication are categorized as non-temporal related clinical data. Figure A illustrates the network architecture of the hybrid artificial neural network (HANN). We used a one-dimension convolution layer with 10 filters to process the temporal related clinical data. For each filter, both length parameters of the convolution window and the convolution stride are set as two, so the layer is simply calculating the inner product between the filter and input. The right side of HANN is a three-layer, fully connected, neural network that deals with the non-temporal related clinical data. A dropout layer with 0.5 dropout rate was applied for reducing overfitting in neural networks. The number of hidden layer neurons (hidden layer_1 and hidden layer_3) was set as 2/3 the size of the input layer, plus the size of the output layer. For the output layers of both right and left sides (dense layer_2 and dense layer_4), the number of neurons was set as two which corresponds to our binary classification problem. The ReLU activation was applied to all hidden layers of HANN. We used a merge layer that sums the output tensors form both network and returns a single tensor as output. Following the merge layer, the final layer is connected to a softmax classifier with dense connections. For the model training, total number of epochs was 150. The mini-batch size was set to 56. We selected the binary cross entropy as the loss function and SGD with learning rate $5e^{-3}$ to be optimizer.

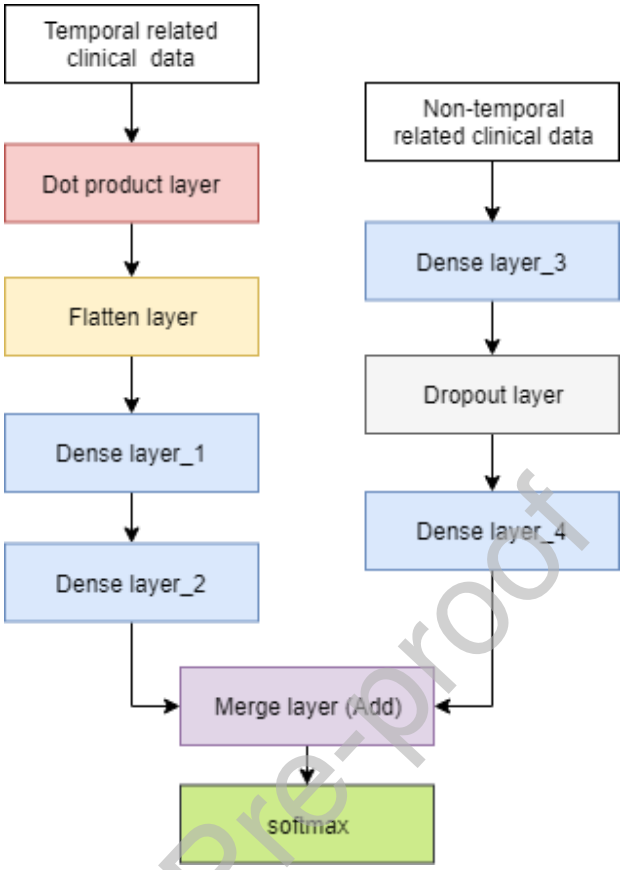
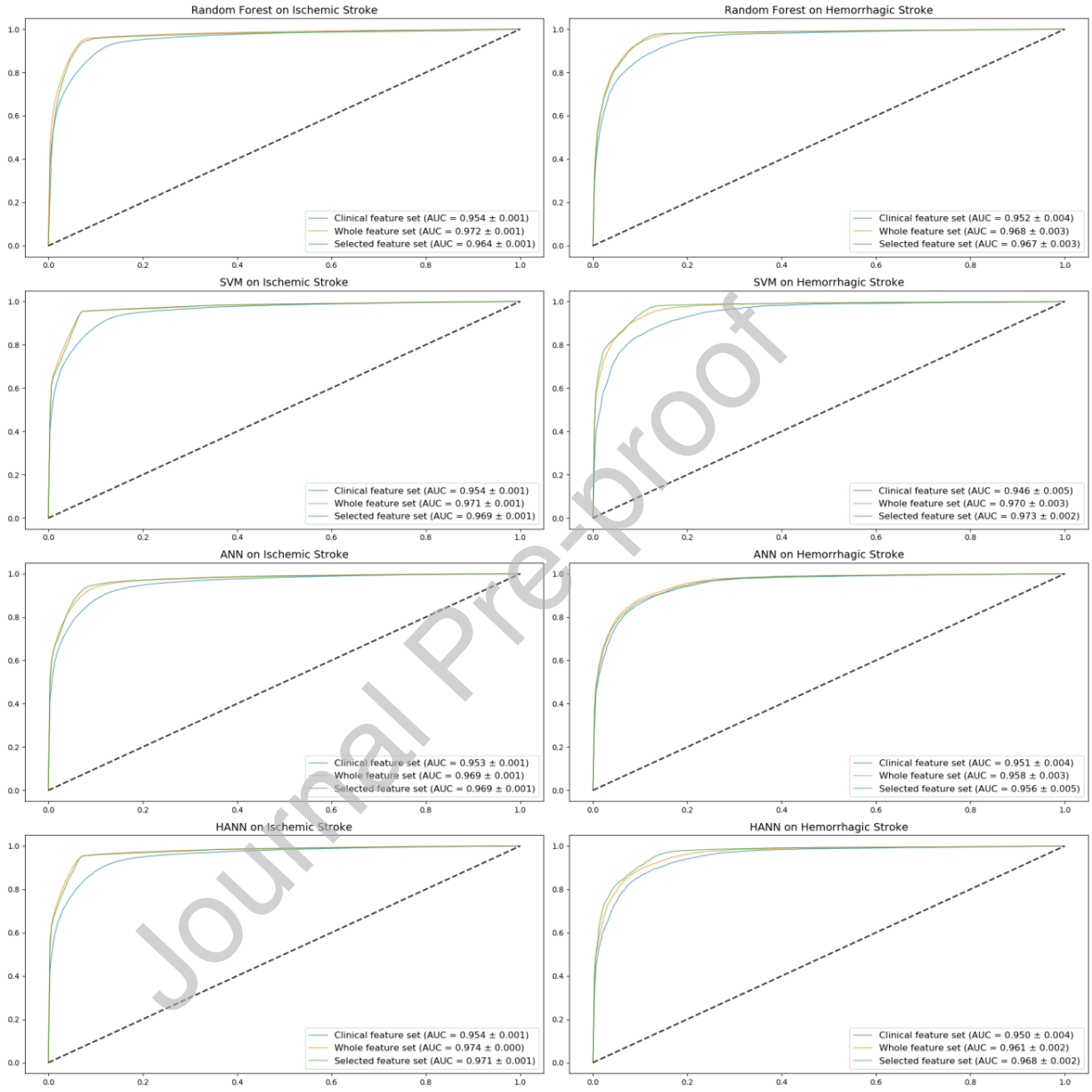
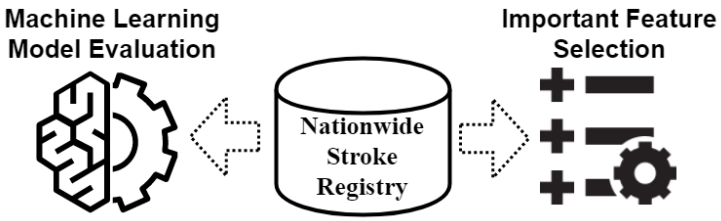


Figure A. The network architecture of the hybrid artificial neural network (HANN)

Appendix Figure B: The receiver operating curve of 90-day stroke outcome prediction models on 10-fold cross-validation training datasets.





Graphical Abstracts