Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients

Miguel Monteiro, Ana Catarina Fonseca, Ana Teresa Freitas, Teresa Pinho e Melo,
Alexandre P. Francisco, José M. Ferro, and Arlindo L. Oliveira

Abstract—Ischemic stroke is a leading cause of disability and death worldwide among adults. The individual prognosis after stroke is extremely dependent on treatment decisions physicians take during the acute phase. In the last five years, several scores such as the ASTRAL, DRAGON, and THRIVE have been proposed as tools to help physicians predict the patient functional outcome after a stroke. These scores are rule-based classifiers that use features available when the patient is admitted to the emergency room. In this paper, we apply machine learning techniques to the problem of predicting the functional outcome of ischemic stroke patients, three months after admission. We show that a pure machine learning approach achieves only a marginally superior Area Under the ROC Curve (AUC) (0.808 ± 0.085) than that of the best score (0.771 ± 0.056) when using the features available at admission. However, we observed that by progressively adding features available at further points in time, we can significantly increase the AUC to a value above 0.90. We conclude that the results obtained validate the use of the scores at the time of admission, but also point to the importance of using more features, which require more advanced methods, when possible.

Index Terms—Ischemic stroke, machine learning, prediction, AUC

1 Introduction and Related Work

Ischemic stroke occurs when there is an obstruction of the blood vessels that supply blood to the brain. This medical condition is one of the leading causes of disability and mortality amongst adults worldwide. Despite advances in treatment, around one-third of patients who survive go on to live with long-term disability [1].

Since stroke treatment is not risk-free, physicians only proceed with treatment when the potential benefits outweigh the perceived risks. For this reason, tools that can help predict the patient's functional outcome using only data available when the patient is admitted to the hospital are extremely useful because they can help inform the treatment decision. In addition, patients and relatives frequently ask questions regarding the individual's prognosis after the stroke and whether the patient will be functionally independent in his or her daily life. The answer to this question is not immediate or straightforward, as a result, any tool that could help determine with high accuracy what the functional outcome of a patient will be at different points in time after an ischemic stroke would be extremely useful.

The most commonly used metric for assessing the functional outcome of a stroke patient is the modified Rankin Scale (mRS) [2]. This 0 to 6 scale measures the degree of disability or dependence of a patient as it relates to daily activities (0 corresponding to full independence and 6 corresponding to death). When trying to predict the patient's functional outcome, it is common to use a binary version of the mRS with only two classes: good and poor outcome. Furthermore, it is also common to measure the mRS 3 months (90 days) after the stroke.

Several efforts have been made by the medical community to create scores that can predict the patient's functional outcome using data readily available at admission. Among them, the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) [3], the DRAGON [4] and the Totaled Health Risks in Vascular Events (THRIVE) [5] scores. These scores are meant to be calculated immediately when the patient is admitted in order to inform the treatment decision.

These scores use statistical analysis to determine the most relevant covariates from a set of pre-selected features by domain experts. Since these scores are meant to be easily calculated by humans using data readily available when the patient is admitted to the emergency service, the resulting statistical models are greatly simplified to create integer-based scores. In practice this means the models' weights are discretized and the number of covariates is artificially reduced, this results in a deterioration of the models' performance. From a machine learning perspective, these scores can be viewed as rule-based classifiers created by domain experts. All the aforementioned scores

- M. Monteiro is with INESC-ID, Lisbon 1000-029, Portugal. E-mail: mab.mtr@gmail.com.
- A.C. Fonseca, T. Pinho e Melo, and J.M. Ferro are with IMM, School of Medicine, University of Lisbon, Lisboa 1649-004, Portugal. E-mail: catarinagfonseca@gmail.com, {tmelo, jmferro}@fm.ul.pt.
- A.T. Freitas, A.P. Francisco, and A.L. Oliveira are with INESC-ID, IST, University of Lisbon, Lisboa 1649-004, Portugal. E-mail: ana.freitas@tecnico.ulisboa.pt, aplf@ist.utl.pt, aml@inesc-id.pt.

Manuscript received 21 Feb. 2018; accepted 26 Feb. 2018. Date of publication 1 Mar. 2018; date of current version 6 Dec. 2018.
(Corresponding author: Miguel Monteiro.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2811471

have been externally validated and report an AUC in the range of 0.70 to 0.80 [6], [7], [8].

Since these scores are all meant to inform treatment, none of them focuses on using data collected after the treatment to determine the functional outcome. In this setting, there is more information available that could be used to give a better prediction and to better inform the physician, patient, and relatives about the patient's functional outcome.

We propose the use of machine learning techniques to predict the patient's functional outcome using information available at different points in time. To the best of our knowledge, there is no report of a pure machine learning approach to the problem of predicting the long-term functional outcome of a stroke patient, nor a comparison between such an approach and the previously mentioned scores. Our approach differs from previous ones in that the prediction is meant to be performed by a computer and in that the model's features are learnt from data instead of selected by domain experts.

The goal of this paper is to use machine learning techniques to predict the functional outcome of a patient three months after the initial stroke. We start by using only the information available at admission to train the classifiers. We then compare the results of the machine learning methods with the results given by the scores, which were designed by domain experts for this specific application. Afterwards, we analyze how the prediction improves as we add more features collected at different points in time after admission. Furthermore, we aim to show how machine learning techniques can be successfully applied to clinical data without losing interpretability of the models, a characteristic which is extremely valuable for medical professionals.

2 Methods

Our original dataset was comprised of 541 patients with acute stroke from the Safe Implementation of Treatments in Stroke (SITS)—Thrombolysis registry [9]. The cohort of patients came from the Hospital of Santa Maria, in Lisbon, Portugal, which is a tertiary university hospital. Even though different hospitals have different cohorts, all hospitals collect the same data for the SITS registry. This registry has baseline data collected at admission, follow-up data collected 2 hours, 24 hours, and 7 days after the initial stroke event and data collected on discharge. In addition, the mRS three months after the event was recorded for 425 patients. All of the patients on the registry were treated using Recombinant Tissue Plasminogen Activator (rtPA).

2.1 Data Description and Pre-Processing

Data cleaning was performed by deleting features that contained only missing values and features which were metadata (e.g.,. record number). In addition, we transformed categorical variables into dummy variables by the means of one-hot-encoding. We converted variables that record times into time differences between variables (e.g., time of the initial event and time of arrival at the hospital becomes time between the event and arrival). We removed patients for which the mRS was not recorded.

After cleaning the data, the resulting dataset had 152 features and 425 patients.

For data imputation, we used the median value for numerical variables (e.g., age) and the mode for discrete variables (e.g., gender), this includes categorical variables. All of the features were scaled to have zero mean and unit variance.

The features obtained from pre-processing could be divided into five sets, depending on the time at which they were collected:

1) Baseline: features collected at admission: the patient's demographic information, past history and risk factors, the time between stroke onset, arrival at the hospital and the start of treatment, the NIH Stroke Scale (NIHSS) [10] (discriminated by field, not just the final result), test results, type of treatment;
2) 2 hours after admission: discriminated NIHSS and test results;
3) Twenty-four hours after admission: discriminated NIHSS, test results, data relating to lesions detected during Computerised Tomography (CT) scan and/or Magnetic Resonance Imaging (MRI), the global outcome as reported by the physician;
4) Seven days after admission: discriminated NIHSS, the global outcome;
5) Discharge: discriminated NIHSS, the global outcome, the treatment the patient was discharged with, the results of cause investigation.

The target variable was the mRS three months after the event. To turn the problem into a binary classification problem, and to compare our results directly with the existing methods, we discretized the mRS into two classes according to:

- Good outcome: defined by $mRS \leq 2$
- Poor outcome: defined by $mRS > 2$

This particular discretization is of medical relevance because it separates the patients who will be able to live a rather normal independent life, from the ones who will require significant assistance.

2.2 Experimental Design

After pre-processing the data, we designed five experiments that aimed to assess with what precision we could predict the patient's mRS three months after admission, given the information available at different points in time.

Table 1 presents the five experiments that were conducted, as well as the total number of patients, the class-split percentages, and the number of features used in each experiment.

Note that the total number of patients and class split percentage varies from experiment to experiment because at each time step we removed the patients who had died between experiments. We made this choice because for patients who had died the mRS at three months would be perfectly defined by the data and immutable over time. This would cause us to overestimate the performance of the proposed methods.

For experiment 1, where only baseline information is available, we also calculated the ASTRAL, DRAGON and THRIVE scores to make possible a direct comparison.

2.3 Scores

To ascertain the relative performance of the proposed machine learning methods, we used three scores as benchmarks. We will not go into detail on how to calculate each

TABLE 1
Experiments

Experiment Number	Feature sets used	# samples	Good outcome	Poor Outcome	# features
1	Baseline	425	51.3% (218)	48.7% (207)	49
2	Baseline, 2h	425	51.3% (218)	48.7% (207)	67
3	Baseline, 2h, 24h	424	51.4% (218)	48.6% (206)	102
4	Baseline, 2h, 24h, 7d	415	52.5% (218)	47.5% (197)	119
5	Baseline, 2h, 24h, 7d, discharge	399	54.6% (218)	45.3% (181)	152

score individually. However, in this section, we give a brief description of the features used by each of the scores. As previously mentioned, all of these scores only use data that is collected when the patient is admitted to the hospital and hence before the treatment.

The ASTRAL score is an integer-based score that takes into account the patient's age, the severity of the stroke through the NIHSS, the time delay from onset to admission, the range of visual defect (NIH_3), the glucose level and the level of consciousness presented by the patient (NIH_1A).

The DRAGON score is a 10-point based score used to predict the 3-month outcome in stroke patients. This score takes into account a dense cerebral artery sign or early infarct signs on the admission CT head scan, the mRS before the stroke, the age, the glucose level, the time delay from onset to treatment and the NIHSS.

Like the DRAGON score the THRIVE score is a 10-point score used to predict the 3-month outcome. However, this score is far more simple, only taking into account the patient's age, the NIHSS, and the previous history of the following chronic diseases: atrial fibrillation, hypertension and diabetes mellitus.

2.4 Classifiers

For our experiments, we used the following classifiers:

- *l1* regularized Logistic Regression;
- Decision Tree;
- Support Vector Machine (SVM);
- Random Forest;
- Xgboost [11].

All of these are commonly used classifiers, and hence we refrain from giving a detailed mathematical explanation of each one.

When analyzing the results, we favour classifiers from which feature importance can be derived. This point is important for physicians in order to able to understand which factors play a role in the patient's recovery.

Except for Xgboost [11], the implementations of the other classifiers can be found in scikit-learn [12].

2.5 Evaluation and Training

To measure the performance of the models we used the AUC [13] since the accuracy is not a good measure of performance when there is a class imbalance. In addition, the AUC is far more informative [14], especially in medical contexts.

To train and validate the model we used 10-fold cross-validation. We present the AUC results as $mean \pm std$ calculated over the 10-folds' validation sets.

In scenarios such as our own where there are not many samples, the choice of the classifier is sometimes irrelevant. Different classifiers will often have very similar performances with differences that are not statistically significant. For this reason, we employed paired t-tests [15] to determine if the differences observed between the AUC of different classifiers were statistically significant.

Considering two classifiers L_A and L_B the algorithm for paired t-tests is presented in Algorithm 1.

Algorithm 1. Paired t-Test [15]

```
1: Partition the available data $D_0$ into $k$ disjoint subsets
   $T_1, T_2, \dots, T_k$ of equal size, where the size is at least 30.
2: For $i$ from 1 to $k$, do
   use $T_i$, for the validation set, and the remaining data for the
   training set $S_i$
     • $S_i \leftarrow D_0 - T_i$
     • $h_a \leftarrow L_A(S_i)$
     • $h_b \leftarrow L_B(S_i)$
     • $\delta_i \leftarrow AUC(h_a) - AUC(h_b)$
3: Return the value $\bar{\delta}$ and $s_{\bar{\delta}}$, where
     $\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$
   and
     $s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\bar{\delta} - \delta_i)^2}$
```

This procedure estimates the difference in AUC between two classifiers, where $\bar{\delta}$ and $s_{\bar{\delta}}$ are the mean and standard deviation that govern this distribution. The t-statistic can the be calculated via Equation (1).

$$t_{stat} = \frac{\bar{\delta}}{s_{\bar{\delta}}}. \tag{1}$$

The number of degrees of freedom that went into producing the sample is given by Equation (2).

$$\nu = k - 1. \tag{2}$$

Knowing the number of degrees of freedom ν and given a certain p-value, the calculated t-statistic can be compared against t-tables to determine if the difference in AUC observed between the two classifiers is statistically significant. For our experiments, we chose a p-value threshold of 0.01. In this case, the sign of $\bar{\delta}$ merely determines which of the classifiers, L_A or L_B, yielded a better AUC.

TABLE 2
AUC for Different Classifiers and Experiments

	Logistic	SVM	Decision Tree	Random Forest	Xgboost	ASTRAL	DRAGON	THRIVE
Exp. 1	0.789 ± 0.091	0.797 ± 0.081	0.758 ± 0.093	$\mathbf{0.808 \pm 0.085}$	0.794 ± 0.084	0.771 ± 0.056	0.751 ± 0.060	0.735 ± 0.055
Exp. 2	0.841 ± 0.070	0.846 ± 0.067	0.803 ± 0.056	0.852 ± 0.071	$\mathbf{0.853 \pm 0.059}$	-	-	-
Exp. 3	0.900 ± 0.063	0.898 ± 0.062	0.877 ± 0.050	0.906 ± 0.063	$\mathbf{0.911 \pm 0.063}$	-	-	-
Exp. 4	$\mathbf{0.926 \pm 0.046}$	0.909 ± 0.057	0.916 ± 0.047	$\mathbf{0.926 \pm 0.041}$	0.925 ± 0.043	-	-	-
Exp. 5	0.924 ± 0.028	0.914 ± 0.039	0.907 ± 0.038	$\mathbf{0.936 \pm 0.034}$	0.930 ± 0.027	-	-	-

To determine the best parameterization for the models we performed a grid search over a set of reasonable values. For Logistic Regression, we searched over the regularization cost parameter. For the SVM we searched over the cost parameter both for the linear and radial kernel. For the Decision Tree, we searched over the maximum number of features considered per split, the minimum number of samples per leaf, the maximum depth of the tree and the minimum number of samples per split. For the Random Forest, we searched over the minimum number of samples per leaf, the minimum number of samples per split and the number of estimators. For the Xgboost classifier, we searched over the learning rate, the maximum depth of the trees, the number of estimators and the sub-sampling rate.

3 Results

Table 2 shows the results of the five experiments for all the classifiers as well as the three scores. Since the scores only use features available at admission, there is no point in recalculating them after experiment 1, as their performance would not change.

From Table 2, we can see that the Random Forest classifier performs best for experiments 1, 3 and 5. Furthermore, we observe that the Xgboost classifier performs best for experiments 2 and 3. Apart from the Decision Tree classifier, we can see that all the classifiers perform better than the scores for experiment 1. Additionally, their performance increases as more time passes from admission and more features are added. For the last two experiments, the best classifiers reach AUC around 0.93. Contrary to this, the performance of the scores is static and does not improve past experiment 1.

TABLE 3
Paired t-Tests Results

Classifier A	Classifier B	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Random Forest	Xgboost					
-	SVM					✓
-	Logistic Regression					
-	Decision Tree	✓	✓			✓
-	ASTRAL		✓	✓	✓	✓
-	DRAGON	✓	✓	✓	✓	✓
-	THRIVE	✓	✓	✓	✓	✓
Xgboost	SVM					
-	Logistic Regression					
-	Decision Tree		✓	✓		✓
-	ASTRAL		✓	✓	✓	✓
-	DRAGON	✓	✓	✓	✓	✓
-	THRIVE		✓	✓	✓	✓
SVM	Logistic Regression					
-	Decision Tree		✓			
-	ASTRAL		✓	✓	✓	✓
-	DRAGON	✓	✓	✓	✓	✓
-	THRIVE		✓	✓	✓	✓
Logistic Regression	Decision Tree		✓			
-	ASTRAL		✓	✓	✓	✓
-	DRAGON		✓	✓	✓	✓
-	THRIVE		✓	✓	✓	✓
Decision Tree	ASTRAL				✓	✓
-	DRAGON			✓	✓	✓
-	THRIVE			✓	✓	✓
ASTRAL	DRAGON					
-	THRIVE					
DRAGON	THRIVE					

Given that the observed performances are so close to each other, it is important to determine if the observed differences are statistically significant. Table 3 shows the results of the paired t-test between all the classifiers and scores for the five experiments. For readability reasons, we omit the numeric value of the t-statistics and instead place a check-mark when the performance of classifier L_A is better than the performance of classifier L_B. The paired t-tests were performed using a p-value of 0.01 and 10-fold cross-validation. The table has been arranged such that the performance of classifier L_B is never superior to one of classifier L_A.

From Table 3, we can see that for experiment 1, for which the scores were designed, the difference in performance between the Random Forest classifier and the DRAGON and THRIVE scores is statistically significant. In addition, for the same experiment, the Xgboost and the SVM classifiers also outperform the DRAGON score. We note that, that the difference in performance between the ASTRAL score and the classifiers for experiment 1 is never statistically significant. It is also relevant to note that after experiment 1, with the exception of the Decision Tree, all classifiers have performances that are statistically significantly better than the scores.

Moreover, we observe that there is no experiment where the differences between the Random Forest, Xgboost and Logistic Regression classifiers are statistically significant. This is not the case for the Decision Tree classifier which often under-performs the other classifiers and the SVM classifier which has a worse performance than the Random Forest classifier for experiment 5.

Due to the large amount of results generated from the five classifiers and the five experiments, we present the Receiver Operating Characteristic (ROC) curves only for the Random Forest classifier. We chose this classifier because it obtained the best overall result, outperforming the most other methods and scores.

Fig. 1 shows the ROC curves of the Random Forest classifier for the five experiments. We only show the mean of the curve not to overcrowd the image with the confidence intervals.

From Fig. 1, we can see that there are two major performance jumps: one between experiments 1 and 2, and another between experiments 2 and 3. Applying paired t-tests between the different experiments for the same classifier (Random Forest), we observed that:

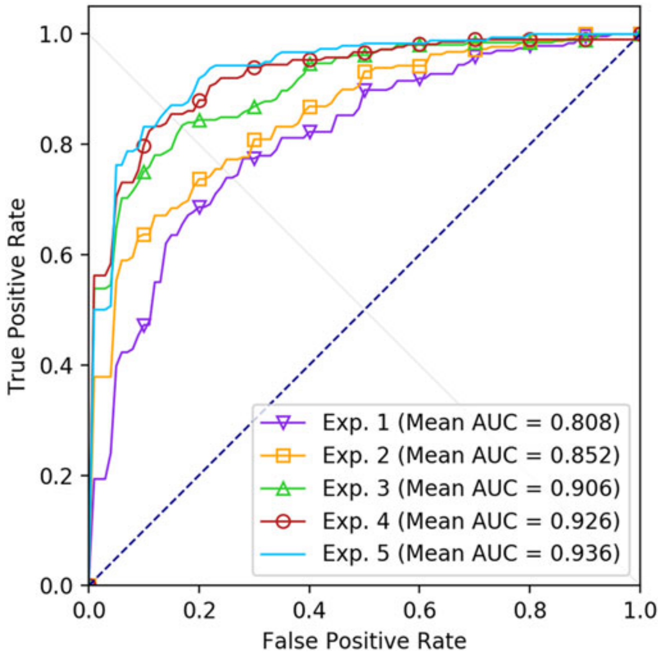

Fig. 1. ROC curves for the random forest classifier.

- All performance increases relative to experiment 1 are statistically significant;
- The performance increase between experiment 5 and 2 is statistically significant;
- All other performance increases were not observed to be statistically significant.

This behaviour was not the same for all classifiers. However, we refrain from presenting the comparison results for all experiments and all classifiers to avoid redundancy. Whilst all of the classifiers showed two performance increases that were statistically significant, their locations in time varied. Nonetheless, the trend was clear: the more time passes after the stroke, the better the prediction and the more likely the performance increase was significant.

Given that the Random Forest, Xgboost and Logistic Regression classifiers have similar performances and all output feature importance, we choose these classifiers to analyze which features are relevant predictors of a poor outcome. In Table 4, we show the top five most important features in the prediction for each of the three classifiers in experiment 1. In Table 5, we show the same information but for experiment 4, when 7 days had already passed from the initial stroke event. By coincidence, all of the features represented in these tables are positively correlated with a predicted poor outcome.

From Table 4, we can see that the classifiers disagree on the order of feature importance. However, there is some common ground between them, for instance, the NIHSS and its sub-component NIH_2, which is related to horizontal extra-ocular movements, always appear in the top five features for all classifiers. In addition, it also interesting to note that features such as age, glucose level, NIHSS, and dense artery sign in CT, which are all known functional outcome predictors that are used by the scores, also appear in the most important features list of the classifiers.

From Table 5, we see that as more time passes from the stroke, the most important features of the three classifiers change and become more homogeneous. At this point in time, the NIHSS (7d), which measures the severity of the symptoms, is the top feature for the three classifiers. The sub-components and previous measurements of the NIHSS also dominate the most important features list for the Random Forest and Logistic Regression classifiers. Interestingly, for the Xgboost classifier, the systolic and diastolic blood pressures, measured at 2 hours after the stroke, occupy the fourth and fifth positions in the most important features list 7 days after the stroke.

4 Discussion

The results presented in Tables 2 and 3 show that some of the scores can be useful when only data collected at admission is available. Even though both the ASTRAL and THRIVE scores achieved a lower AUC than most machine learning methods, the paired t-tests for the ASTRAL score never had a worse performance than for any of the classifiers, and the THRIVE score was only bested by the Random Forest classifier. On the other hand, the DRAGON score showed a worse performance than the Random Forest,

TABLE 4
Feature Importance (Baseline)

Rank	Random Forest	Xgboost	Logistic Regression
1	NIHSS	NIHSS	NIH_5A
2	NIH_2	Systolic blood pressure	NIH_2
3	Glucose (mg)	Age	Dense artery sign in CT
4	Systolic blood pressure	Glucose (mg)	Congestive heart failure (risk factor)
5	Age	NIH_2	NIHSS

TABLE 5
Feature Importance (7 Days)

Rank	Random Forest	Xgboost	Logistic Regression
1	NIHSS (7d)	NIHSS (7d)	NIHSS (7d)
2	NIHSS (24h)	Age	NIH_5A (7d)
3	NIHSS (2h)	NIHSS (24h)	NIH_5B (7d)
4	NIH_5A (7d)	Systolic blood pressure (2h)	Age
5	NIH_4 (7d)	Diastolic blood pressure (2h)	NIH_5B (24h)

Xgboost and SVM classifiers. This indicates that this score is the weakest of the three scores under analysis.

Regardless, we speculate that if more data were available to train the classifiers, we would observe more scores having statistically worse performances than the classifiers. Training the classifiers with more data could mean an increase in the average AUC, which would certainly not happen for the scores since they are not being learnt from data.

Looking at Table 2 and at Fig. 1 it becomes clear that as more time passes from the initial stroke event, and the more features are added to the classifiers, the more their performance increases. This performance increase is already statistically significant when using features collected 2 hours after the stroke (experiment 2). Even though these predictions made later in time cannot be used to inform treatment, they can still be used to inform physicians, patients and relatives about the recovery prospects of the patient. This information can be used to take the necessary precautions and improve the quality of living of stroke victims. In addition, we can see that for the later predictions the AUC of the best classifiers is around 0.93. This value is already good enough to have a very reliable prediction of the functional outcome of the patient.

The observed steady improvement of the predictions with time is explained by the fact that the symptoms of stroke are maximal on onset and decrease in severity with time. As a result, the state of the patient after 7 days or on discharge is much less likely to change significantly than immediately after the stroke or at 2 hours after. As more time passes the more certain we can be about the prediction because the patient's state is less likely to change.

By looking at Table 4, we can see that different classifiers disagree on feature importance. Nonetheless, the NIHSS and its sub-components play a major role in the prediction. This is to be expected since these metrics measure the severity of the symptoms, and the more severe the symptoms the less likely recovery is. Age at the time of development of a cerebral lesion has also been previously described as an important determinant of outcome. Younger patients tend to recover more easily and completely than older patients [16]. Other top features such as the glucose level, infarct signs on the admission CT head scan, systolic blood pressure (hypertension) are all known stroke outcome predictors that are also used by the scores.

From Table 5, we see that the features that were important predictors when the patient was admitted change over time. At this point in time, the features that become more important are the NIHSS and its sub-components measured at different points in time. The NIH_5A and NIH_5B which are associated with motor arm strength are top predictors. There is biological plausibility for the selection of these variables since arm strength is essential to the performance of most daily activities that are evaluated by the mRS. Age and blood pressure (hypertension) are also top predictors.

5 Conclusion and Future Work

We conclude that machine learning techniques can be effectively used to predict the functional outcome of an ischemic stroke patient three months after the initial event. The resulting AUC can range up to 0.936 depending on the classifier used and on the point in time at which the prediction is made. Furthermore, we have validated the use of scores when only data at admission is available and have shown that some machine learning models can be interpreted to derive new knowledge.

In the future we wish to improve our prediction by using more and richer records, both by using more cohorts from the SITS database and by using our own database which we are currently developing with more complex data. Moreover, we aim to incorporate the use of image and genetic information and to take advantage of the longitudinal aspect of the data.

Acknowledgments

This work was supported by national funds through the Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and by project PRECISE - accelerating progress toward the new era of precision medicine (PAS application C443089117-00042059).

References

[1] T. Truelsen, B. Piechowski-Jóźwiak, R. Bonita, C. Mathers, J. Bogousslavsky, and G. Boysen, "Stroke incidence and prevalence in Europe: A review of available data," *Eur. J. Neurology : The Official J. Eur. Fed. Neurological Societies*, vol. 13, no. 6, pp. 581–98, 2006.

[2] J. Rankin, "Cerebral vascular accidents in patients over the age of 60. II. Prognosis.," *Scottish Med. J.*, vol. 2, pp. 200–215, May 1957.

[3] G. Ntaios, M. Faouzi, W. Ferrari, J Lang, K. Vemmos, and P. Michel, "An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL score," *Neurology*, vol. 78, no. 2, pp. 1916–22, 2012.

[4] D. Strbian, A. Meretoja, F. J. Ahlhelm, J. Pitkäniemi, P. Lyrer, M. Kaste, S. Engelter, and T. Tatlisumak, "Predicting outcome of IV thrombolysis - Treated ischemic stroke patients: The DRAGON score," *Neurology*, vol. 78, no. 6, pp. 427–432, 2012.

[5] A. C. Flint, S. P. Cullen, B. S. Faigeles, and V. A. Rao, "Predicting long-term outcome after endovascular stroke treatment: The totaled health risks in vascular events score," *Amer. J. Neuroradiology*, vol. 31, no. 7, pp. 1192–1196, 2010.

[6] C. Cooray, M. Mazya, M. Bottai, L. Dorado, O. Skoda, D. Toni, G. A. Ford, N. Wahlgren, and N. Ahmed, "External validation of the ASTRAL and DRAGON scores for prediction of functional outcome in stroke," *Stroke*, vol. 47, no. 6, pp. 1493–1499, 2016.

[7] A. C. Flint, B. S. Faigeles, S. P. Cullen, H. Kamel, V. A. Rao, R. Gupta, W. S. Smith, P. M. Bath, and G. A. Donnan, "Thrive score predicts ischemic stroke outcomes and thrombolytic hemorrhage risk in vista," *Stroke*, vol. 44, no. 12, pp. 3365–3369, 2013.

[8] P. Gaucher and T. Hildncr, "Totaled health risks in vascular events score predicts clinical outcome and symptomatic intracranial hemorrhage in chinese patients after thrombolysis,"*Stroke*, vol. 18, no. 6, 2015, Art. no. 11.

[9] N. Wahlgren, N. Ahmed, A. Dávalos, G. A. Ford, M. Grond, W. Hacke, M. G. Hennerici, M. Kaste, S. Kuelkens, V. Larrue, K. R. Lees, R. O. Roine, L. Soinne, D. Toni, and G. Vanhooren, "Thrombolysis with alteplase for acute ischaemic stroke in the safe implementation of thrombolysis in stroke-monitoring study (SITS-MOST): An observational study," *Lancet*, vol. 369, no. 9558, pp. 275–282, 2007.

[10] T. Brott, H. P. Adams, C. P. Olinger, J. R. Marler, W. G. Barsan, J. Biller, J. Spilker, R. Holleran, R. Eberle, and V. Hertzberg, "Measurements of acute cerebral infarction: A clinical examination scale," *Stroke*, vol. 20, no. 7, pp. 864–870, 1989.

[11] T. Chen and C. Guestrin, "XGBoost," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[13] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *J. Math. Psychology*, vol. 12, no. 4, pp. 387–415, 1975.

[14] A. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[15] T. Mitchell, *Machine Learning* New York, NY, USA: McGraw-Hill Education, 1997.

[16] R. L. Harvey, "Predictors of functional outcome following stroke," *Phys. Med. Rehabil. Clinics North*, vol. 26, pp. 583–598, 2015.

Miguel Monteiro received the bachelor's degree in electrical engineering and computer science from the University of Lisbon and the double master's degree in electrical engineering and computer science from the University of Lisbon and the Royal Institute of Technology, Stockholm (KTH), in 2016. He is currently a researcher in INESC-ID where he works in machine learning applied to ischemic stroke.

Ana Catarina Fonseca received three master's degree in stroke medicine from the University Danube Krems, Austria, in neuroscience from the University of Lisbon, and in public health from Harvard University, in 2014, and the PhD degree in medicine from the Faculty of Medicine, University of Lisbon, within the medical specialty of neurology, in 2014. She is a researcher in the Instituto de Medicina Molecular, teaches pharmacology and neurology at the Faculty of Medicine-University of Lisbon, and practices medicine at the Hospital de Santa Maria. She is currently the vice-president of the Portuguese Neurological Society and a active member of the European Stroke Organization (member of the Education and Membership committee, secretary of the stroke group of the value of treatment project of the European Brain Council). Her main research interests include cryptogenic stroke, heart-brain interactions, and biomarkers.

Ana Teresa Freitas is a full professor in the Department of Computer Science and Engineering, Instituto Superior Técnico, University of Lisbon. Her main scientific expertise is on the areas of algorithms and data mining, bioinformatics, human genetics, and health informatics. She is also the CEO and co-founder of the startup company Heart-Genetics, Genetics, and Biotechnology SA.

Teresa Pinho e Melo received the MD degree from the University of Lisbon. She is a licensed neurologist and part of the teaching staff of Neurology Department of the University of Lisbon. From 1990 to 1991 she worked as a "Médicin Assistent Boursier" from the Neurological Center of "Centre Hospitalier Universitaire Vaudois – Lausanne". Currently she is the coordinator of the stroke unit in Hospital de Santa Maria, Lisbon and the SITS national coordinator of Portugal and SITS Hospital de Santa Maria coordinator.

Alexandre P. Francisco received the PhD degree in computer science and engineering. He is currently an assistant professor in the CSE Department, IST, Universidade de Lisboa. His current research interests include algorithmics, computational science, graphs, and programming, with applications on network mining, and large data processing.

José M. Ferro received the graduate degree in 1975 and the PhD degree in medicine from the the University of Lisbon, in 1987. He was a postdoctoral fellow in the Department of Clinical Neurological Sciences, London, Canada. He is currently a full professor of neurology and president of the "Conselho de Escola" of the School of Medicine, University of Lisbon. He is the director of the Department of Neurosciences and Mental Health and director of the neurology service in the Hospital de Santa Maria, Centro Hospitalar Lisboa Norte. He is the head of the José Ferro Lab in the Instituto de Medicina Molecular, University of Lisbon. He was president of the European Neurological Society and was recently elected for the Board of the Word Stroke Organization. He is also a member of the Scientific Panel for Health of the Directorate-General for Research and Innovation, Directorate E–Health, of the European Commission. He has authored or co-authored 309 papers published in international journals and 60 book chapters. His main area of research interest is cerebrovascular disease, with a focus on cerebral venous thrombosis, cryptogenic stroke, and cognitive and psychiatric consequences of stroke.

Arlindo L. Oliveira received the BSc and MSc degrees in electrical and computer engineering from Lisbon Technical University, and the PhD degree in electrical engineering and computer science from the University of California, Berkeley, in 1986, 1989, and 1994, respectively. He is currently a professor in Instituto Superior Técnico. He is also a senior researcher with INESC-ID. His research interests include bioinformatics, systems biology, string processing, algorithm design, combinatorial optimization, machine learning, logic synthesis, and automata theory. He is a senior member of the IEEE.