



OPEN

## Random forest-based prediction of stroke outcome

Carlos Fernandez-Lozano<sup>1,2</sup>, Pablo Hervella<sup>3</sup>, Virginia Mato-Abad<sup>4</sup>, Manuel Rodríguez-Yáñez<sup>5</sup>, Sonia Suárez-Garaboa<sup>4</sup>, Iria López-Dequidt<sup>5</sup>, Ana Estany-Gestal<sup>6</sup>, Tomás Sobrino<sup>3</sup>, Francisco Campos<sup>3</sup>, José Castillo<sup>3</sup>, Santiago Rodríguez-Yáñez<sup>4✉</sup> & Ramón Iglesias-Rey<sup>3✉</sup>

We research into the clinical, biochemical and neuroimaging factors associated with the outcome of stroke patients to generate a predictive model using machine learning techniques for prediction of mortality and morbidity 3-months after admission. The dataset consisted of patients with ischemic stroke (IS) and non-traumatic intracerebral hemorrhage (ICH) admitted to Stroke Unit of a European Tertiary Hospital prospectively registered. We identified the main variables for machine learning Random Forest (RF), generating a predictive model that can estimate patient mortality/morbidity according to the following groups: (1) IS + ICH, (2) IS, and (3) ICH. A total of 6022 patients were included: 4922 (mean age  $71.9 \pm 13.8$  years) with IS and 1100 (mean age  $73.3 \pm 13.1$  years) with ICH. NIHSS at 24, 48 h and axillary temperature at admission were the most important variables to consider for evolution of patients at 3-months. IS + ICH group was the most stable for mortality prediction [ $0.904 \pm 0.025$  of area under the receiver operating characteristics curve (AUC)]. IS group presented similar results, although variability between experiments was slightly higher ( $0.909 \pm 0.032$  of AUC). ICH group was the one in which RF had more problems to make adequate predictions ( $0.9837$  vs.  $0.7104$  of AUC). There were no major differences between IS and IS + ICH groups according to morbidity prediction ( $0.738$  and  $0.755$  of AUC) but, after checking normality with a Shapiro Wilk test with the null hypothesis that the data follow a normal distribution, it was rejected with  $W = 0.93546$  ( $p$ -value  $< 2.2e-16$ ). Conditions required for a parametric test do not hold, and we performed a paired Wilcoxon Test assuming the null hypothesis that all the groups have the same performance. The null hypothesis was rejected with a value  $< 2.2e-16$ , so there are statistical differences between IS and ICH groups. In conclusion, machine learning algorithms RF can be effectively used in stroke patients for long-term outcome prediction of mortality and morbidity.

Stroke is the second leading cause of death and the third leading cause of disability in the world. Approximately 15 million people will experience a stroke episode every year worldwide of which 33% is left with a permanent disability, whereas 40% of the cases will result in death, and by 2030 will result in the annual loss of over 200 million (death or disability) globally<sup>1</sup>. Developing an appropriate long-term management plan and studying the progress in the management of stroke patients is necessary in order to organize healthcare structures for the coming years<sup>1-5</sup>. Predicting functional outcome after stroke would help clinicians make patient-specific decisions<sup>6-10</sup>.

Machine learning (ML) provides a promising tool for disease evolution prediction and it is being increasingly used in biomedical studies. The application of ML in healthcare is widely anticipated as a key step towards improving care quality, and would play a fundamental role in the development of learning healthcare systems. Large scale studies in the general literature provide evidence in favor of some classifier families such as Random

<sup>1</sup>Department of Computer Science and Information Technologies, Faculty of Computer Science, CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain. <sup>2</sup>Grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos. Imagen Médica y Diagnóstico Radiológico (RNASA-IMEDIR). Instituto de Investigación Biomédica de A Coruña (INIBIC). Complejo Hospitalario Universitario de A Coruña (CHUAC), SERGAS, Universidade da Coruña, A Coruña, Spain. <sup>3</sup>Clinical Neurosciences Research Laboratory (LINC), Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain. <sup>4</sup>Software Engineering Laboratory, Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, Campus de Elviña, 15071 A Coruña, Spain. <sup>5</sup>Stroke Unit, Department of Neurology, Health Research Institute of Santiago de Compostela (IDIS), Hospital Clínico Universitario, Rúa Travesa da Choupana, s/n, 15706 Santiago de Compostela, Spain. <sup>6</sup>Unit of Methodology of the Research, Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain. ✉email: santiago.rodriguez@udc.es; ramon.iglesias.rey@sergas.es

	IS + ICH (n = 6022)	IS (n = 4922)	ICH (n = 1100)
<b>Demographic variables</b>			
Age (years)	72.1 ± 13.7	71.9 ± 13.8	73.3 ± 13.1
Female gender (%)	44.5	44.8	41.5
Arterial hypertension (%)	66.7	63.7	60.7
Diabetes mellitus (%)	23.4	24.1	20.4
Alcohol consumption (%)	12.2	11.5	15.4
Smoking (%)	15.4	16.4	10.7
Dyslipidemia (%)	35.4	35.1	36.7
Peripheral arterial disease (%)	5.7	5.9	4.6
Ischemic heart disease (%)	10.8	11.3	8.6
Atrial fibrillation (%)	20.8	24.1	18.1
Previous transient ischemic attack (%)	5.4	6.1	2.5
Previous ischemic stroke (%)	13.1	13.6	9.8
Previous intracerebral hemorrhage, %	2.4	0.9	9.8
Previous anticoagulants (%)	9.5	8.5	14.1
Previous platelet antiaggregants (%)	22.9	24.4	16.5

**Table 1.** Demographic variables of the experimented dataset of patients summarized by group.

Forest (RF) in terms of classification performance<sup>11,12</sup>. RF has recently been used successfully in a wide range of biomedical applications, such as the automatic detection of pulse during electrocardiogram-based cardiopulmonary resuscitation or in breast cancer diagnosis using mammography images<sup>13–18</sup>.

As to stroke, most studies focus on the use of ML methods to detect ischemic stroke (IS) lesions using neuroimaging data<sup>19–23</sup> and outcome estimation<sup>24–28</sup>. It has only been recently, however, that a study evaluated stroke outcome prediction at 3 months also in a group of non-traumatic intracerebral hemorrhage (ICH) patients using a nationwide disease registry<sup>27</sup>. Previous studies concluded that ML techniques can be effective to predict functional outcome of IS long-term patients or for prediction of symptomatic intracranial haemorrhage following thrombolysis from CT images. However, all works agree on the need to carry out further studies in order to confirm results, incorporate new variables and resolve their limitations or weaknesses.

Taking into account the prevalence of cerebrovascular diseases, accurately predicting stroke evolution is essential to stratify the rehabilitation care that should be administered, especially to patients with the best chance of recovery. The administration of rehabilitative therapies to those who are unlikely to benefit from them is inefficient for the Healthcare System and inconvenient and unproductive for patients. A predictive model that identifies stroke patients at risk of deterioration would make it possible to select/follow-up patients for reperfusion treatments, and increase the control of therapeutic homeostasis, thus addressing the needs of each patient individually. Furthermore, regarding to new regenerative cellular or molecular therapies, it is essential to identify the most suitable patients to respond accurately to treatments. We hypothesized that models developed with ML techniques based on the demographic, clinical, biochemical and neuroimaging variables obtained in the first 48 h after stroke are accurate stroke mortality and morbidity predictors at 3 months.

## Results

We included in the study 6022 patients; 4922 (81.8%) presented with IS and 1100 (18.2%) with ICH. We excluded 228 patients, who died during the first 24 h, and 84 with no follow-up at 3 months. The 65 features of the different groups included in the experimented dataset are shown in Tables 1, 2 and 3. Figure 1 lists flowchart of patient groups with their functional outcome and divided into morbidity and mortality.

Of the 4922 IS patients for this study, 55.2% were male and 44.8% female; the mean age was 71.9 ± 13.8 years. According to the TOAST classification, 1127 patients were classified as atherothrombotic (22.9%), 1786 as cardioembolic (36.3%), 428 as lacunar (8.7%) and 1520 as undetermined (30.9%). Poor functional outcome at 3 months was found in 47.5% of IS patients, thus showing a morbidity of 33.4% and a mortality of 13.2%.

Of the 1100 ICH patients, 58.5% were male and 41.5% female; the mean age was 73.3 ± 13.1 years. ICH etiology was related with 506 hypertensive patients (46%), 114 with amyloid angiopathy (10.4%), with 156 anticoagulants (14.2%) and 323 with other causes (29.4%). 58.6% of ICH patients showed poor outcome at 3 months, with a morbidity of 27.6% and a mortality of 30.2%.

Using the filter feature selection, the dataset was reduced to only 7 variables: National Institute of Health Stroke Scale score at admission [NIHSS (0)]; NIHSS score at 24 h [NIHSS (24)]; NIHSS score at 48 h [NIHSS (48)]; Axillary temperature at admission [T(0)]; Early neurological deterioration [ED]; Leukocytes at admission [LEU (0)]; and blood glucose at admission [GLU (0)] as with these variables, RF was much more stable and deviations or variations between experiments could be reduced.

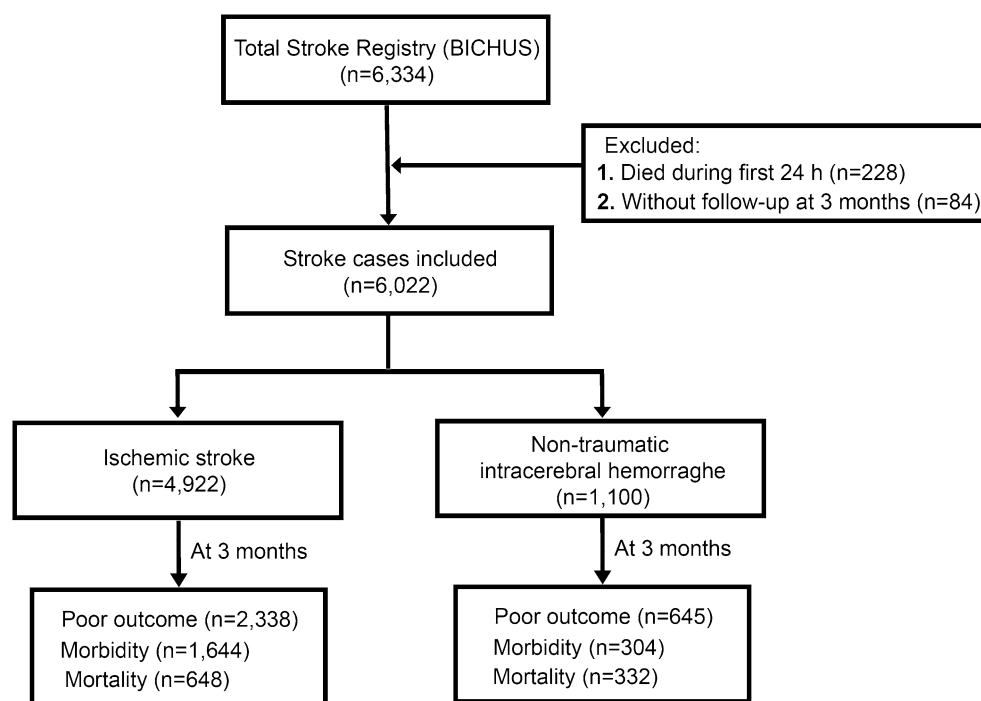
	IS + ICH (n = 6022)	IS (n = 4922)	ICH (n = 1100)
<b>Clinical/Neuroimaging variables</b>			
Stroke on awakening (%)	8.3	9.1	4.6
Previous mRS	0 [0, 1]	0 [0, 1]	1 [0, 1]
Time from stroke onset, minutes	239.1 ± 175.2	240.8 ± 167.4	231.3 ± 206.1
NIHSS score at admission	13 [7, 19]	13 [8, 19]	13 [7, 18]
NIHSS score at 24 h	8 [13, 16]	7 [3, 15]	12 [6, 19]
NIHSS score at 48 h	7 [2, 15]	6 [2, 14]	12 [4, 20]
Early neurological deterioration (%)	7.7	5.8	16.5
<b>TOAST</b>			
Atherothrombotic (%)	–	22.9	–
Cardioembolic (%)	–	36.3	–
Lacunar (%)	–	8.7	–
Undetermined (%)	–	30.9	–
Others (%)	–	1.2	–
Intravenous fibrinolysis (%)	–	22.7	–
Thrombectomy (%)	–	5.2	–
DWI at admission (ml)	–	33.3 ± 76.9	–
TC volume 4th–7th day (ml)	–	51.1 ± 82.3	–
<b>Hemorrhagic transformation of IS</b>			
IH1 (%)	–	7.0	–
IH2 (%)	–	3.1	–
PH1 (%)	–	1.7	–
PH2 (%)	–	1.2	–
<b>Etiology of ICH</b>			
Hypertensive (%)	–	–	46.0
Amyloid (%)	–	–	10.4
Anticoagulants (%)	–	–	14.2
Others/Undetermined (%)	–	–	29.4
Hematoma volume at admission (ml)	–	–	40.3 ± 46.2
Hematoma volume 4th–7th day (ml)	–	–	51.9 ± 48.1
Total hematoma volume (ml)	–	–	68.3 ± 53.1
Volume of hypodensity (ml)	–	–	15.2 ± 17.9
Hematoma growth (ml)	–	–	11.9 ± 27.6
<b>Topography</b>			
Deep hemispherics (%)	–	–	50.0
Lobar (%)	–	–	39.6
Cerebellar (%)	–	–	4.7
Breinstem (%)	–	–	3.8
Primary intraventricular (%)	–	–	1.9
Axillary temperature at admission (°C)	36.4 ± 0.7	36.4 ± 0.7	36.6 ± 0.8
Blood glucose at admission (mg/dl)	137.6 ± 56.3	137.3 ± 57.9	138.9 ± 48.1
Sedimentation rate (mm)	26.4 ± 23.1	26.5 ± 23.1	26.2 ± 23.1
Glycosylated hemoglobin (%)	6.1 ± 2.1	6.1 ± 2.3	5.8 ± 0.9
LDL cholesterol (mg/dl)	101.9 ± 42.9	112.5 ± 44.4	109.6 ± 35.2
HDL cholesterol (mg/dl)	41.2 ± 18.5	41.8 ± 18.5	38.8 ± 18.3
Triglycerides (mg/dl)	118.3 ± 63.1	121.2 ± 65.1	109.4 ± 50.7
Platelets (× 10 <sup>3</sup> /ml)	215.4 ± 82.9	217.7 ± 83.7	203.3 ± 77.9
Hemoglobin (g/dl)	13.7 ± 1.9	13.8 ± 1.9	13.5 ± 2.1
DBP at admission (mmHg)	81.9 ± 16.1	81.5 ± 15.8	84.3 ± 17.2
SBP at admission (mmHg)	152.9 ± 27.3	152.5 ± 27.3	155.5 ± 27.4

**Table 2.** Clinical and neuroimaging variables of the experimented dataset of patients summarized by group.

**Prediction of mortality.** Figure 2A shows the most important variables for the model associated to IS, ICH or IS + ICH patient groups in relation to the mortality prediction. The value shown is, in all cases, the sum of the

	IS + ICH (n = 6022)	IS (n = 4922)	ICH (n = 1100)
<b>Molecular markers</b>			
Leukocytes at admission ( $\times 10^3/\text{ml}$ )	8.9 $\pm$ 3.1	9.1 $\pm$ 3.2	8.8 $\pm$ 3.3
Fibrinogen at admission (mg/dl)	443.9 $\pm$ 101.7	444.5 $\pm$ 101.8	444.1 $\pm$ 101.5
C-reactive protein admission (mg/dl)	2.7 $\pm$ 3.8	3.6 $\pm$ 4.2	5.2 $\pm$ 5.2
Microalbuminuria (mg/24 h)	7.9 $\pm$ 26.2	5.9 $\pm$ 25.9	16.7 $\pm$ 30.0
NT-pro-BNP levels (pg/ml)	915.9 $\pm$ 1563.7	1581.2 $\pm$ 1886.1	1013.8 $\pm$ 3620.2
<b>Outcome at 3 months</b>			
mRS	2 [1, 4]	2 [0, 4]	3 [1, 6]
Poor outcome (%)	49.6	47.5	58.6
Morbidity (%)	35.0	33.4	27.6
Mortality (%)	16.3	13.2	30.2

**Table 3.** Molecular markers and outcome at 3 months of the experimented dataset of patients summarized by group.



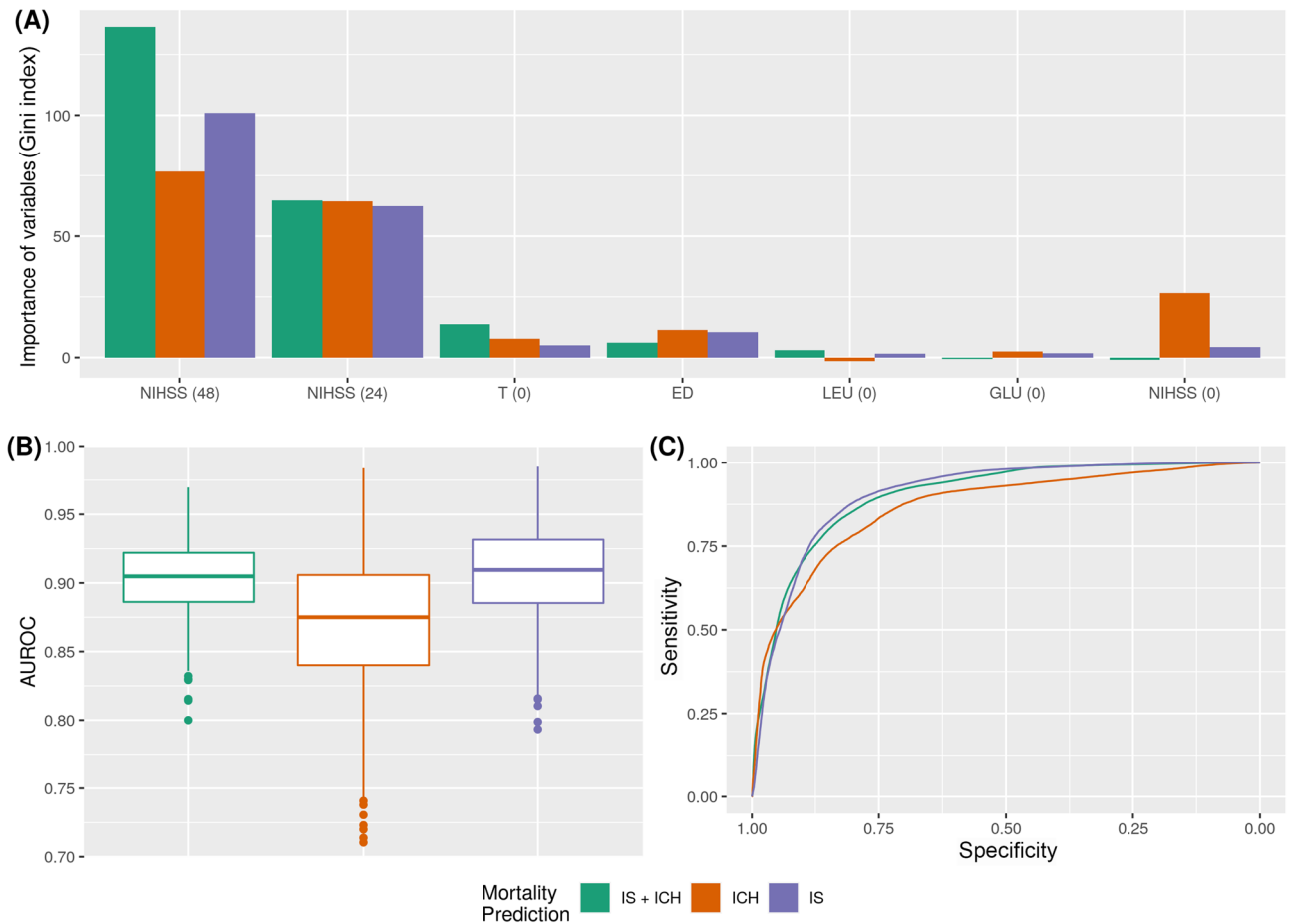
**Figure 1.** Flowchart of patient groups and functional outcome.

importance obtained by the algorithm for the variable in each of the experiments internally.

The most important variables, taking into account the three groups of patients analyzed, were NIHSS (48) and NIHSS (24). In the IS and ICH patient groups, the importance of ED, T (0) and NIHSS (0) should be highlighted. NIHSS (0) was also observed to be more important in patients with ICH than in those with IS when the models do not have data from both types of patients. It seems, however, that its importance is significantly reduced when the model has the complete set. Finally, LEU (0) and GLU (0) are variables that help balance the results for the complete model, reducing the variability of the individual IS or ICH models among all variables.

The variation obtained between all the repetitions performed in area under the receiver operating characteristics (ROC) curve (AUC) terms is detailed in Fig. 2B,C for the three experiments performed. The complete problem with two types of patients is the most stable with a minimum deviation between experiments (median of  $0.904 \pm 0.025$  of AUC and  $0.825 \pm 0.030$  of accuracy (ACC)). On the other hand, the ICH problem is the one in which RF has more problems to make adequate predictions as the range of results varies in more than 20 AUC points between the best ( $0.9837$  of AUC with  $0.94$  of ACC) and the worst experiment ( $0.7104$  of AUC and  $0.6122$  of ACC) and values for 100 repetitions of  $0.875 \pm 0.048$  of AUC and  $0.8 \pm 0.052$  of ACC. This prediction is therefore the most complex for the model.

As to the IS problem, RF presented similar values to those of the IS-ICH prediction problem although variability between experiments is slightly higher ( $0.909 \pm 0.032$  of AUC and  $0.833 \pm 0.040$  of ACC). This led us to



**Figure 2.** Mortality prediction for IS + ICH, IS and ICH groups. **(A)** Main variables for the machine learning model: NIHSS score at admission [NIHSS (0)]; NIHSS score at 24 h [NIHSS (24)]; NIHSS score at 48 h [NIHSS (48)]; Axillary temperature at admission [T(0)]; Early neurological deterioration [ED]; Leukocytes at admission [LEU (0)]; and Blood glucose at admission [GLU (0)]. **(B)** AUROC values obtained. **(C)** ROC curves for the Random Forest classifier.

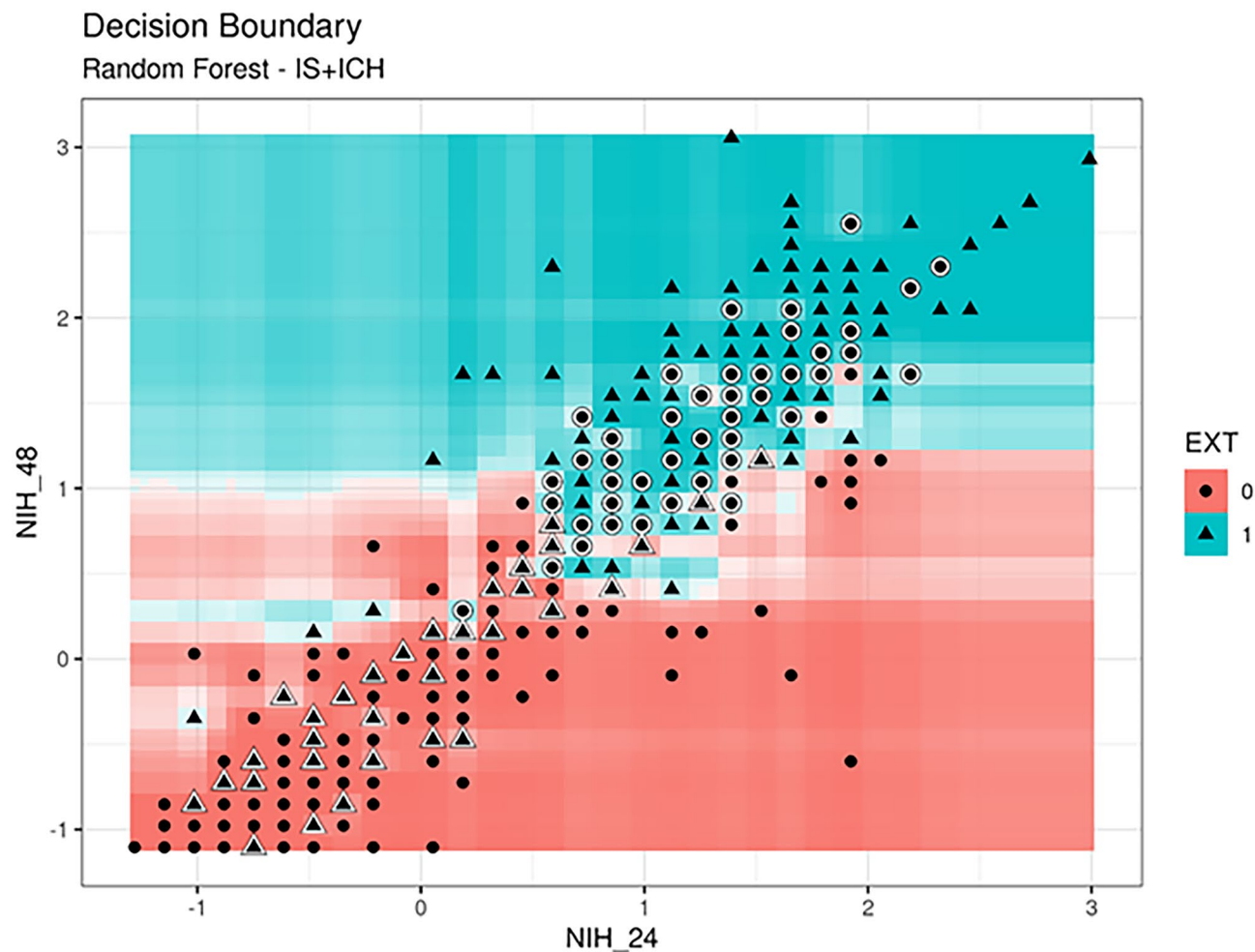
conclude that there is enough information within the selected variables so that when RF has enough patients in the dataset, the model predicts very accurately which patients are most likely to die on the basis of the data collected at admission. It was also observed that when there is a greater amount of data and patients are stratified into the three categories, the model is much more stable, and results are better. A 2D-heatmap of mortality predictions against NIHSS (48) and NIHSS (24) is detailed in Fig. 3 in order to explain the decision boundary of the model. Note that the misclassified items are highlighted and that the intensity of the colors also indicates the certainty of the prediction. We showed the IS + ICH group as it was the most stable for mortality prediction ( $0.904 \pm 0.025$  of AUC).

**Prediction of morbidity.** Figure 4A shows that NIHSS (48) and NIHSS (24) are again the most important variables for the model associated to the three groups studied in relation to morbidity prediction. However, it seems that the variables ED for the IS and total groups, and GLU (0) for ICH patients provide relevant predictive capacity. NIHSS (0) was identified by the model as a variable with negative effects on classification for the IS patient group as it worsens prediction. For the ICH patient group, however, NIHSS (0) is a key variable.

Figure 4B,C shows that there were no major differences between the IS and IS + ICH groups (0.738 and 0.755 of AUC and 0.683 and 0.700 of ACC) but with the ICH (0.667 of AUC and 0.618 of ACC).

**AUC with 7 input variables.** Figure 5 shows the comparison of ROC curves for the most important variables of the ML model associated to the groups IS + ICH, IS or ICH patients in relation to the mortality and morbidity prediction. IS + ICH and IS patient groups curves revealed NIHSS (24) and NIHSS (48) variables with the best AUC values obtained for both morbidity [0.677 (CI 95% 0.662–0.692) vs. 0.703 (CI 95% 0.686–0.719)]; and 0.669 (CI 95% 0.654–0.684) vs. 0.697 (CI 95% 0.680–0.713)] and mortality [0.888 (CI 95% 0.876–0.900) vs. 0.892 (CI 95% 0.878–0.906); and 0.897 (CI 95% 0.885–0.908) vs. 0.899 (CI 95% 0.885–0.913)] prediction.

The ICH group presented NIHSS (0) and NIHSS (24) with the best AUC values obtained for morbidity [0.591 (CI 95% 0.553–0.629) and 0.588 (CI 95% 0.550–0.626)]; and NIHSS (24) and NIHSS (48) for mortality [0.865 (CI 95% 0.838–0.891) and 0.873 (CI 95% 0.847–0.899)] estimation.



**Figure 3.** 2D-heatmap of mortality (EXT) predictions against NIHSS(48) and NIHSS(24). Model results are shown for the IS + ICH group, as it was the most stable for mortality prediction ( $0.904 \pm 0.025$  of AUC). Red areas correspond to patients who do not die (0), blue areas correspond to patients who die (1), and misclassified items are highlighted.

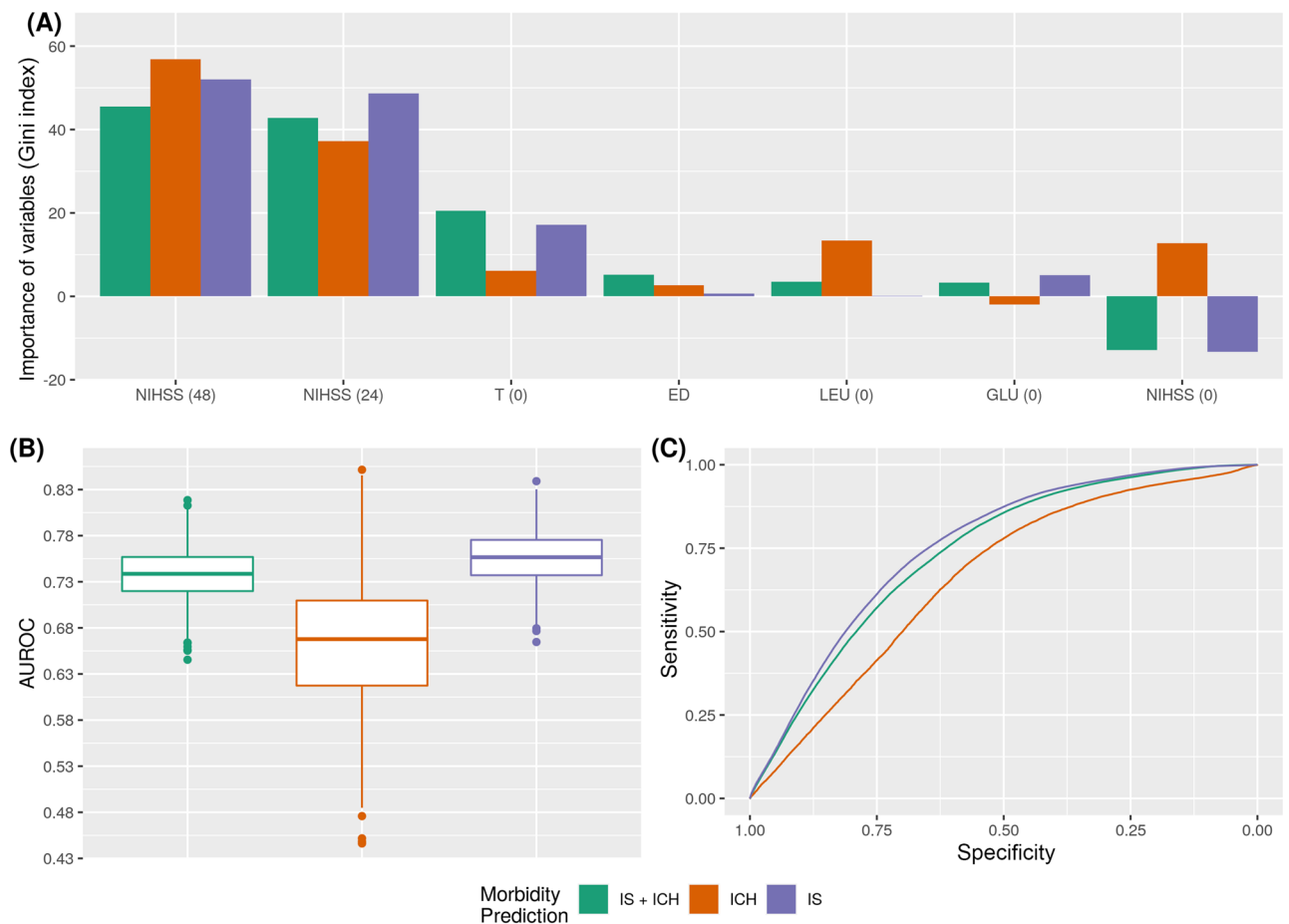
## Discussion

It is difficult but essential to accurately predict functional outcomes after stroke. Outcome prediction plays an important role in long-term decision making, patient treatment, organization of Health Centers, and domestic conditions. Plans could be developed on the basis of a better prediction of the degree of recovery of each patient with appropriate and individualized rehabilitation measures that take into account the domestic and economic conditions, leading to shared decisions with patients, relatives, and sociomedical centers<sup>27,29</sup>.

The conventional approach to the evaluation of stroke outcomes data resorts to classical statistical models (logistic regression). Logistic regression models identify and validate predictive variables. Their main advantage is that they can be easily implemented and interpreted<sup>24</sup>. ML algorithms have the potential to outperform conventional regression because they are able to capture nonlinearities and complex interactions among multiple predictor variables. They can also handle large-scale multi-institutional data, with the added advantage of easily incorporating newly available data to improve prediction performance, and a better handling of a large number of predictors<sup>30</sup>.

A recent systematic review found that ML was not superior to logistic regression in clinical prediction modeling<sup>31</sup>. However, inconsistent conclusions have been often found when comparing the performance of classical models to different machine learning algorithms in clinical applied studies. These studies agree that further research is needed to assess the feasibility and acceptance of ML applications in clinical practice<sup>32,33</sup>. Different research works have proposed strategies for stroke prediction based on ML algorithms with excellent results, but with a great diversity in the variables analyzed (clinical, molecular markers or imaging), calibration/training protocols performed, and models implemented (neural networks, tree-based and kernel-based methods). Some limitations of these previous studies are due to: (1) the low sample size used, (2) the characteristics of the patients evaluated; most of the studies only evaluate IS sub-groups, such as, IS patients treated with rTPA or endovascular intervention, (3) studies used demographic, clinical, molecular or neuroimaging variables independently and uncorrelated, and (4) the small number of variables used in ML models. Asadi et al. performed dichotomized modified Rankin Scale (mRS) models of acute ischemic stroke ( $n = 107$ ) and presented 0.6 AUC of ANN and 70%

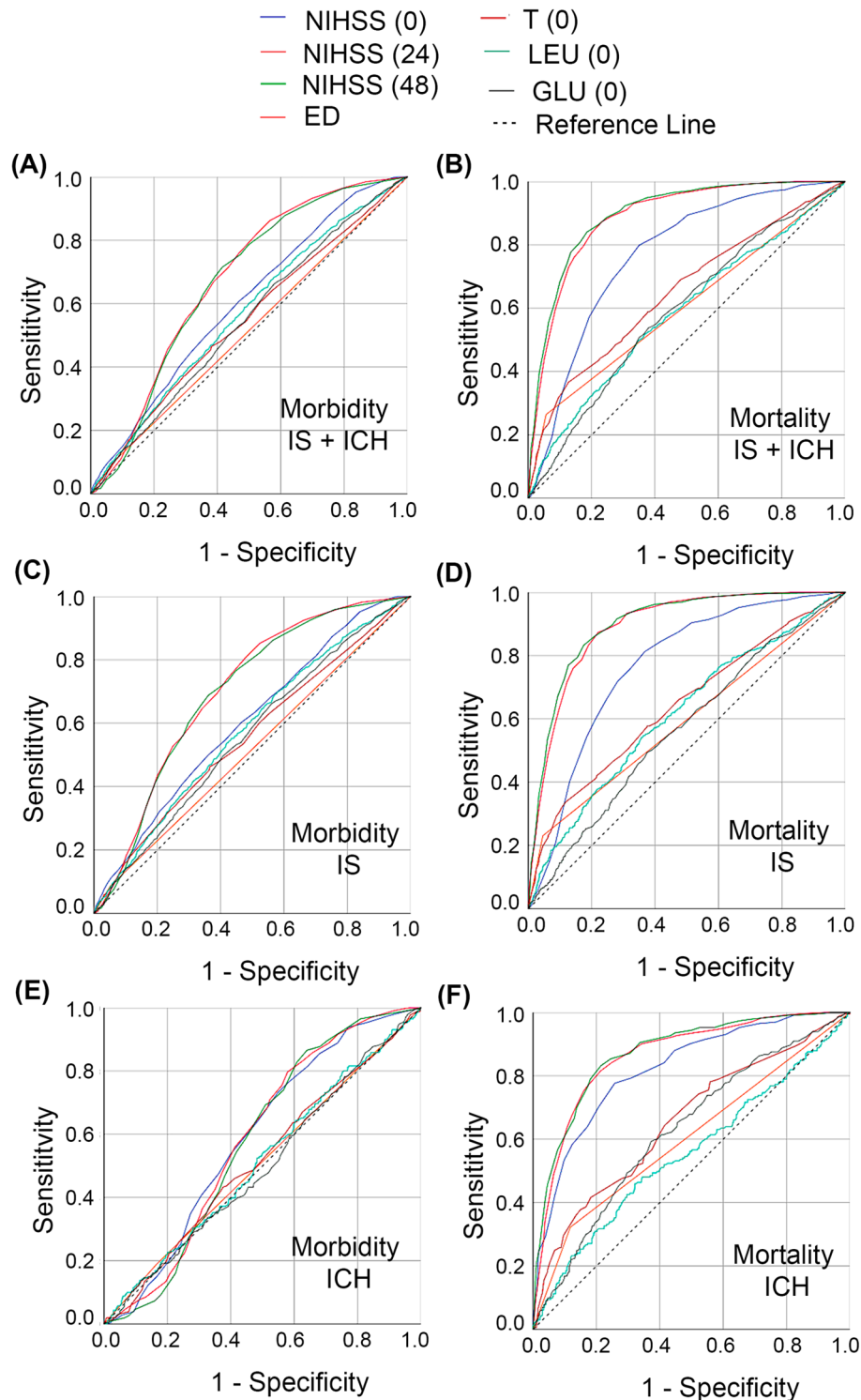




**Figure 4.** Morbidity prediction for IS + ICH, IS and ICH groups. **(A)** Main variables for the machine learning model: NIHSS score at admission [NIHSS (0)]; NIHSS score at 24 h [NIHSS (24)]; NIHSS score at 48 h [NIHSS (48)]; Axillary temperature at admission [T(0)]; Early neurological deterioration [ED]; Leukocytes at admission [LEU (0)]; and Blood glucose at admission [GLU (0)]. **(B)** AUROC values obtained. **(C)** ROC curves for the Random Forest classifier.

accuracy of SVM<sup>24</sup>. Bentley et al. developed an SVM model to study acute ischemic stroke patients ( $n = 116$ ) at risk for symptomatic intracranial hemorrhage using CT brain images<sup>23</sup>. Monterio et al. applied ML techniques (RF, Xgboost, SVM and Decision tree) to predict the functional outcome of ischemic stroke patients ( $n = 425$ ) treated with Recombinant Tissue Plasminogen Activator (rtPA) 3 months after the initial stroke. They started using only the information available at admission and then they went on to analyze how prediction improves by adding more features collected at different points in time after admission. The ML approach achieved AUC of 0.808 when using the features available at admission and as new features were progressively added, AUC increased to a value above 0.90<sup>25</sup>. Heo et al. researched into the applicability of machine learning-based models with a prospective cohort of 2604 patients with acute ischemic stroke. The AUC obtained was 0.888 for DNN model, 0.857 for RF model, and 0.849 for logistic regression model<sup>26</sup>. Recently, Lin et al. used a Taiwan Stroke Registry ( $n = 40,293$ ) to evaluate several ML approaches (SVM, RF, ANN, and HANN) for 90-day stroke outcomes prediction. ML techniques presented over 0.94 AUC in both ischemic and hemorrhagic stroke using preadmission and inpatient data<sup>27</sup>. Alaka et al. concluded that both logistic regression and ML models had comparable predictive accuracy at 90 days when validated internally (AUC range = [0.65–0.72]) and externally (AUC range = [0.66–0.71]) in acute IS patients after endovascular treatment ( $n = 614$ – $684$ )<sup>28</sup>.

In our study, we analyzed a ML model of stroke prediction at 3 months using the Hospital's Stroke Registry (BICHUS) on the basis of demographic, clinical, molecular and neuroimaging variables. Mortality and morbidity were evaluated by identifying the main variables for the ML model. Data were studied as a whole (IS + ICH) or as independent subsets. Our ML classifiers exhibited high performance with over 0.90 AUC in the three groups evaluated in relation to the mortality outcome. The IS group had the best results ( $n = 4922$ ). The model indicates that the most relevant variables are NIHSS (48) and NIHSS (24). In addition, the variable NIHSS (0) is also important for the ICH patients ( $n = 1100$ ). The rest of the variables provide information marginally, although the importance of T(0) and ED should not be disregarded. On the other hand, AUC over 0.75 was found in the three groups evaluated in relation to the morbidity outcome. The model developed indicates that the most relevant variables are NIHSS (48) and NIHSS (24), although ED for the IS group and GLU (0) for the ICH



**Figure 5.** Comparison of ROC curves of 7 variables selected for machine learning experiments for mortality and morbidity prediction at 3 months of the different patient groups evaluated. (A,B) Morbidity and mortality of IS + ICH group. (C,D) Morbidity and mortality of IS group. (E–F) Morbidity and mortality of ICH group.

provide predictive capability. Compared to ROC curve analysis of the 7 input variables, ML classifier has a high performance in three groups, with the NIHSS (0), NIHSS (24) or NIHSS (48) as the most influential predictors.

The findings in this report are subject to at least four limitations. First, this was a retrospective, single-center study with a relatively small clinical dataset. The intrinsic need for large training datasets may affect the accuracy of ML models in studies that could be overadjusted by irrelevant clinical predictors, or some predictors may be underestimated, thus increasing random errors. It is important to note that the variables selected in our study



had been previously identified by means of a T-test and supervised by expert neurologists. However, in the future, both training and validation procedures will need to include a multicenter dataset and a prospective study to verify the model and the variables obtained. Second, the IS and the ICH patient groups were unbalanced. We consider, however, that the two types of stroke should be studied independently to find both differences and similarities. Third, we used RF machine learning algorithms, although other models like DNN or deep logistic regression could be used for comparison purposes. Fourth, we used typical clinical variables as inputs for the ML model, and we did not stratify patients in different subgroups, which could improve the results presented. We consider that it would be useful to evaluate the common variables from a clinical point of view, so once again emphasis is on the importance of NIHSS, axillary temperature and blood glucose. The major strengths of the present study include the large sample size (6022 patients; 4922 with IS and 1100 with ICH), which enabled study of the combination of different stroke types in detail (IS + ICH, IS, and ICH). Furthermore, to derive a global risk score for stroke, we have evaluated/interrelated demographic, clinical, biochemical and neuroimaging variables. Another distinctive feature of this analysis compared with previous studies is that we also included molecular markers associated with inflammation (leukocytes, fibrinogen and C-reactive protein), endothelial and atrial dysfunction (microalbuminuria and NT-proBNP).

## Conclusions

Machine learning algorithms, particularly Random Forest, can be effectively used in long-term outcome prediction of mortality and morbidity of stroke patients. NIHSS at 24, 48 h and axillary temperature are the most important variables to consider in the evolution of the patients at 3 months. Future studies could incorporate the use of imaging and genetic information. Furthermore, the robust model developed could be used in other applications and different scopes with similar data; such as traumatic brain injury, or dementia (Alzheimer's and Parkinson's disease).

## Materials and methods

**Patient selection.** The dataset used in this research work consisted of patients with IS and ICH admitted to the Stroke Unit of the Hospital Clínico Universitario of Santiago de Compostela (Spain), who were prospectively registered in an approved data bank (BICHUS). All patients were treated by a certified neurologist according to national and international guidelines. Exclusion criteria for this analysis were: (1) patients who died during the first 24 h, and (2) loss of follow-up (personal interview or telephone contact) at 3 months.

The analysis of the data for this study was retrospective, from September 2007 to September 2017. This research was carried out in accordance with the Declaration of Helsinki of the World Medical Association (2008) and approved by the Ethics Committee of Santiago de Compostela (2019/616). All patients or their relatives signed the informed consent for inclusion in the registry and for anonymous use of their personal data for research purposes.

**Demographic, clinical, molecular and neuroimaging variables.** The registry includes demographic variables, previous medical history and vital signs. Blood samples for hemogram, biochemistry and coagulation tests were obtained and analyzed at the central hospital laboratory. Neurological deficit was evaluated by a certified neurologist using the National Institute of Health Stroke Scale (NIHSS) at admission, and every 24/48 h during hospitalization. The modified Rankin Scale (mRS) was used to evaluate functional outcome at discharge and at 3 months<sup>33,34</sup>.

Effective reperfusion of IS patients was defined as  $\leq 8$  points in the NIHSS during the first 24 h. Early neurological deterioration was defined as  $\geq 4$  points in NIHSS within the first 48 h with respect to baseline NIHSS score. Poor functional outcome was defined as  $mRS > 2$  at 3 months, morbidity as  $3 \leq mRS \leq 5$ , and mortality as  $mRS = 6$ . Ischemic stroke diagnosis was made using the TOAST criteria<sup>35</sup>.

Computed Tomography (CT) was performed in all patients and Magnetic Resonance Imaging (MRI) in selected patients at admission. Follow-up CT scan after fibrinolysis or thrombectomy was performed in all IS patients at 24 h, and CT at 48 h or when neurological deterioration was detected and between the 4th–7th day. ICH and perihematomal edema volumes were calculated using the ABC/2 method<sup>36</sup>. ICH topography was classified as lobar when it predominantly affected the cortical/subcortical white matter of the cerebral lobes or as deep when it was limited to the internal capsule, the basal ganglia or the thalamus. All neuroimaging tests were analyzed by a neuro-radiologist supervised by the above certified neurologist.

**Outcome endpoints.** The objective of this research work was to identify the main predictors for the machine learning model in order to generate a predictive model using machine learning techniques for the prediction of mortality and morbidity of stroke patients according to their stratification to one of the following groups: 1) IS + ICH, 2) IS, and 3) ICH.

**Machine learning.** We used the RF algorithm for the prediction of mortality and morbidity of stroke patients. RF is an ensemble learning method, i.e., a strategy that aggregates many predictions to reduce the variance and improve the robustness and precision of outputs<sup>37–39</sup>. A remarkable characteristic of the RF is that it provides an internal measure of the relative importance of each feature on the prediction. This model generally works very well for any type of problem, regardless of size and even if the data are unbalanced or missing<sup>37</sup>. It also makes it possible to analyze the importance of the variables used by the model. To this end, the Gini importance index was calculated. This index measures the increase in impurity of each variable in the model when selected in the random distribution process. Each time a node selects a variable, the Gini impurity index for the two child

nodes is lower than in the parent. It is not a simple final summation of the values obtained in all the trees for each variable but a weighting.

Generally speaking, all ML algorithms have a number of hyperparameters that must be optimized to obtain the best results for the particular problem they are analyzing. We used R<sup>40,41</sup> and the following packages: mlr<sup>42</sup> to calculate the best number of trees (ranging from 500 to 1000); Random Forest<sup>43</sup> for our experiments; and ggplot2<sup>44</sup> graphics for data analysis.

**Data pre-processing.** Balanced classes in classification problems are critical for ML algorithms. When analyzing the problem, we initially obtained prediction values in AUC lower than 0.65 in the best of cases using 65 features and four different ML algorithms, which is considered a bad performance value for prediction. This is mainly due to the fact that in our dataset there is a high percentage of patients who survived versus patients who died, and we found noise and correlation between features, confounding the predictors. These numbers show an unbalanced problem that needs to be addressed since a predictive model for patient death is being generated.

The following are the two main approaches to balance the data: (1) oversampling the minority class or (2) undersampling the class where the data has more examples. Although these are very powerful techniques that are able to increase the performance of the classifiers, they must be handled cautiously, more so in medical problems, to prevent overadjustments or the loss of generalization capacity in the models when new synthetic samples are included (oversampling) at the learning phase of the algorithms<sup>45</sup>. In this work, undersampling methods (random undersampling from the majority class) were assessed for class balancing purposes. To ensure that the undersampling process is fair and that the generalization capability of the models is not biased we ran 100 repetitions, each with a different random undersampling, of a tenfold cross-validation experiment to observe the behavior and the stability of results. The more stability the better the random removing process. This means that the remaining samples of the majority class captured the underlying knowledge of the class.

The experimental design developed to analyze the original data included: a data preprocessing phase and the balancing of the subclasses; a tenfold cross-validation and 100 repetitions. For each of these repetitions, the position of each patient in the dataset was randomized. The preprocessed data were also randomized to avoid any potential process-related bias.

The problem was broken down into six different but complementary and informative problems: mortality and morbidity prediction with IS, ICH or IS + ICH patients. This approach sought to analyze more exhaustively the differences between the different types of patients when predicting death/poor outcome and to analyze whether the variables with more weight in the prediction were the same in all the cases.

In order to identify which of the 65 variables available are the most informative, we performed feature selection. There are mainly three different approaches for feature selection in machine learning: filter, wrapper and embedded<sup>46</sup>. Filter methods assess the relevance of each feature by looking only at the intrinsic properties of the data (independent of the algorithms). We calculated a feature relevance score (T-test) on the training data, and low-scoring features were removed choosing a manual cut-off point to reduce the variance of the models (see Supplementary Fig. S1 and Supplementary Material).

**Statistical analysis.** For the descriptive study of the quantitative variables, results were expressed as percentages for categorical variables and as mean (SD) or median (quartiles) for the continuous variables, depending on whether their distribution was normal or not. The Kolmogorov–Smirnov test was used for testing the normality of the distribution. To measure the performance of the model we used the area under the receiver operating characteristics (ROC) curve (AUC or AUROC)<sup>47</sup>. To train and validate the model we used tenfold cross validation. AUC results are presented as mean ± SD calculated over the tenfold validation sets. To test whether an AUC of logistic regression and ML models prediction could obtain similar results, ROC curve analysis was used to compare the 7-input variables selected for ML experiments of the different patient groups as potential morbidity and mortality clinical markers at 3 months. The statistical descriptive analysis was conducted in SPSS 25.0 (IBM, Chicago, IL) for Mac.

## Data availability

All data are available within the text of the manuscript. Further anonymized data could be made available to qualified investigators upon reasonable request.

Received: 24 December 2020; Accepted: 26 April 2021

Published online: 12 May 2021

## References

1. Neuhaus, A. A. *et al.* Neuroprotection in stroke: The importance of collaboration and reproducibility. *Brain* **140**, 2079–2092 (2017).
2. Bramlett, H. M. & Dietrich, W. D. Pathophysiology of cerebral ischemia and brain trauma: Similarities and differences. *J. Cereb. Blood Flow Metab.* **24**, 133–150 (2004).
3. Burns, J. D., Fisher, J. L. & Cervantes-Arslanian, A. M. Recent advances in the acute management of intracerebral hemorrhage. *Neurosurg. Clin. N. Am.* **29**, 263–272 (2018).
4. Béjot, Y. *et al.* Epidemiology of stroke in Europe and trends for the 21st century. *Presse. Med.* **45**, e391–e439 (2016).
5. Rodríguez-Castro, E. *et al.* Trends in stroke outcome in the last ten years in a European tertiary hospital. *BMC Neurol.* **18**, 164 (2018).
6. Fens, M. *et al.* Multidisciplinary care for stroke patients living in the community: A systematic review. *J. Rehabil. Med.* **45**, 321–330 (2013).
7. Sen, A. *et al.* Continuous hemodynamic monitoring in acute stroke: An exploratory analysis. *Emerg. Med.* **15**, 345–350 (2014).
8. Elmarazy, A. *et al.* Desmoteplase for acute ischemic stroke: A systematic review and metaanalysis of randomized controlled trials. *CNS Neurol. Disord. Drug Targets* **16**, 789–799 (2017).

9. Lin, Y. *et al.* Endovascular thrombectomy for acute ischemic stroke: A meta-analysis. *CNS Neurol. Disord. Drug Targets* **314**, 1832–1843 (2015).
10. Baratloo, A. *et al.* Effects of telestroke on thrombolysis times and outcomes: A meta-analysis. *Prehosp. Emerg. Care* **22**, 472–484 (2018).
11. Olson, R. S. *et al.* Data-driven advice for applying machine learning to bioinformatics problems. *Pac. Symp. Biocomput.* **23**, 192–203 (2018).
12. Deist, T. M. *et al.* Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Med. Phys.* **46**, 1080–1087 (2019).
13. Elola, A. *et al.* ECG-based pulse detection during cardiac arrest using random forest classifier. *Med. Biol. Eng. Comput.* **57**, 453–462 (2019).
14. Shaikhina, T. *et al.* Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed. Signal Process. Control* **52**, 456–462 (2019).
15. Liu, Y. *et al.* Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. *Expert. Syst. Appl.* **72**, 306–316 (2017).
16. Chowdhury, A. R., Chatterjee, T. & Banerjee, S. A random forest classifier-based approach in the detection of abnormalities in the retina. *Med. Biol. Eng. Comput.* **57**, 193–203 (2019).
17. Dmitriev, K. *et al.* Classification of pancreatic cysts in computed tomography images using a random forest and convolutional neural network ensemble. *Med. Image Comput. Assist. Interv.* **10435**, 150–158 (2017).
18. Alickovic, E. & Subasi, A. Breast cancer diagnosis using GA feature selection and rotation forest. *Neural Comput. Appl.* **28**, 753–763 (2017).
19. Mitra, J. *et al.* Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *Neuroimage* **98**, 324–335 (2014).
20. Maier, O. *et al.* Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. *J. Neurosci. Methods* **240**, 89–100 (2015).
21. McKinley, R. *et al.* Fully automated stroke tissue estimation using random forest classifiers (FASTER). *J. Cereb. Blood Flow Metab.* **37**, 2728–2741 (2017).
22. Subudhi, A. *et al.* Automated approach for detection of ischemic stroke using Delaunay Triangulation in brain MRI images. *Comput. Biol. Med.* **103**, 116–129 (2018).
23. Bentley, P. *et al.* Prediction of stroke thrombolysis outcome using CT brain machine learning. *NeuroImage Clin.* **4**, 635–640 (2014).
24. Asadi, H. *et al.* Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS ONE* **9**, e88225 (2014).
25. Monteiro, M. *et al.* Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 1953–1959 (2018).
26. Heo, J. *et al.* Machine learning-based model can predict stroke outcome. *Stroke* **49**, A194A (2018).
27. Lin, C. H. *et al.* Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Comput. Methods Programs Biomed.* **190**, 105381 (2020).
28. Alaka, S. A. *et al.* Functional outcome prediction in ischemic stroke: A comparison of machine learning algorithms and regression models. *Front. Neurol.* **11**, 889 (2020).
29. Stinear, C. M. Prediction of recovery of motor function after stroke. *Lancet Neurol.* **9**, 1228–1232 (2010).
30. Alpaydin E. Introduction to machine learning. Cambridge, Massachusetts (2010)
31. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
32. Lynam, A. L., Dennis, J. M. & Owen, K. R. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: Application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn. Progn. Res.* **4**, 6 (2020).
33. Montaner, J. & Álvarez-Sabín, J. NIHSS Stroke Scale and its adaptation to Spanish. *Neurologia* **21**, 192–202 (2006).
34. Bonita, R. & Beaglehole, R. Modification of rankin scale: recovery of motor function after stroke. *Stroke* **19**, 1497–1500 (1998).
35. Adams, H. P. Jr. *et al.* Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST Trial of org 10172 in acute stroke treatment. *Stroke* **24**, 35–41 (1993).
36. Sims, J. R. *et al.* ABC/2 for rapid clinical estimate of infarct, perfusion, and mismatch volumes. *Neurology* **72**, 2104–2110 (2009).
37. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
38. Breiman, L. Random forests. *Mach. Learn.* **24**, 123–140 (1996).
39. Amit, Y. & Geman, D. Shape quantization and recognition with randomized trees. *Neural. Comput.* **9**, 1545–1588 (1997).
40. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria (2020). <https://www.R-project.org/>.
41. R Core Team. The R Project for Statistical Computing. Vienna, Austria (2019).
42. Bischl, B. *et al.* mlr: machine learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016).
43. Liaw, A. & Wiener, M. Classification and regression by random forest. *R News* **2**, 18–22 (2002).
44. Wickham H. ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag, New York. ISBN: 978-3-319-24277-4 (2016)).
45. Lopez, V. *et al.* An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **250**, 113–141 (2013).
46. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007).
47. Bradley, A. A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **7**, 1145–1159 (1997).

## Acknowledgements

None.

## Author contributions

Drs. J.C., S.R.-Y., and R.I.-R. conceived the scientific idea, designed the experiments, assisted with statistical analysis, and provided input in the writing of the manuscript. Drs. C.F.-L., V.M.-A., and S.S.-G. performed machine learning analysis, and provided input in the writing of the manuscript. Drs. M.R.-Y., and I.L.-D., had major role in the acquisition of data. Neuroimaging study, interpreted the data, revised the manuscript. A.E.-G., T.S., F.C., P.H. provided discussions on the project throughout, interpreted data and also provided input in the writing of the manuscript. All authors reviewed the manuscript.

## Funding

This study was partially supported by grants from the Spanish Ministry of Science and Innovation (SAF2017-84267-R), Xunta de Galicia (Axencia Galega de Innovación (GAIN): IN607A2018/3), Instituto de Salud Carlos III (ISCIII) (PI17/00540, PI17/01103), Spanish Research Network on Cerebrovascular Diseases RETICS-INVIC-TUS PLUS (RD16/0019) and by the European Union FEDER program. T. Sobrino (CPII17/00027), F. Campos (CPII19/00020) are recipients of research contracts from the Miguel Servet Program (Instituto de Salud Carlos III). General Directorate of Culture, Education and University Management of Xunta de Galicia (ED431G/01,252 ED431D 2017/16), "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23), Competitive Reference Groups (ED431C 2018/49), Spanish Ministry of Economy and Competitiveness via funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503), European Regional Development Funds (FEDER).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-89434-7>.

**Correspondence** and requests for materials should be addressed to S.R.-Y. or R.I.-R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021