

# Predicting Mortality in Patients with Stroke Using Data Mining Techniques

Zahra Hadianfard <sup>1</sup>, Hadi Lotfnezhad Afshar <sup>1</sup> , Surena Nazarbaghi <sup>2</sup> , Bahlol Rahimi <sup>1</sup> ,  
Toomas Timpka <sup>3</sup> 

<sup>1</sup> Department of Health Information Technology, School of Allied Medical Sciences, Urmia University of Medical Sciences, Urmia, Iran

<sup>2</sup> Department of Neurology, School of Medicine, Urmia University of Medical Sciences, Urmia, Iran

<sup>3</sup> Department of Health, Medicine, and Caring Sciences, Linköping University, Linköping, Sweden

Corresponding author: Hadi Lotfnezhad Afshar (lotfnezhadafshar.h@umsu.ac.ir)

## Abstract

The mortality due to stroke is increasing. Accurate prediction of stroke-caused death is very important for healthcare. Data mining methods are novel ways to predict these mortality risks. The aim of this study is to employ popular data mining algorithms to predict the survival of stroke patients and extract decision rules. The data on stroke patients (n=4149) were collected from paper medical records. Missing data were managed using the multiple imputation method. Also, the target variable was balanced using methods such as over-sampling, under-sampling and Synthetic Minority Oversampling (SMOTE). The support vector machine (SVM), decision tree, and logistic regression (LR) algorithms were employed to predict the survival of stroke patients. Also, the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm was used to extract the decision rules from the main dataset. LR outperformed other algorithms in terms of accuracy (76.96%), sensitivity (79.06%) and kappa (33.34). However, specificity (65.35%) and AUC (0.77) scores were lower than those of other algorithms. An independent dataset with 234 records was selected to challenge the LR algorithm with the best performance from the main dataset. After employing this algorithm on the external validation dataset, its performance was improved in accuracy (79.91%), sensitivity (83.94%), kappa (39.26) and AUC (0.8), but not in specificity (60.98%). The constructed model predicted the survival of stroke patients with high scores and useful rules were extracted for clinical usage.

## Keywords

Data mining; Decision trees; Stroke; Survival; Logistic regression; Iran.

**Citation:** Hadianfard, Z., Afshar, H. Z., Nazarbaghi, S., Rahimi, B., & Timpka, T. (2022). Predicting Mortality in Patients with Stroke Using Data Mining Techniques. *Acta Informatica Pragensia*, 11(1), 36–47. <https://doi.org/10.18267/j.aip.163>

**Academic Editor:** Zdenek Smutny, Prague University of Economics and Business, Czech Republic

**Copyright:** © 2022 by the author(s). Licensee Prague University of Economics and Business, Czech Republic.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

# 1 Introduction

While stroke was the fifth common cause of death worldwide in 1990, it was ranked third in 2017. In recent decades, the incidence rates have decreased by 42% in developed countries but increased by more than 100% in low- and middle-income countries (Gebreyohannes et al., 2019). For instance, the incidence of stroke in Iran has been estimated to span up to 140 cases in every 100,000 people (Azarpazhooh et al., 2010; Hosseini et al., 2010). Ischemic stroke is sustained by the highest number of patients (80%) with different types of stroke (Boehme et al., 2017). Almost 60% of all new cases of ischemic stroke occur in people under 70 years (Lindsay et al., 2019). However, incidence of this disease is approximately one decade earlier in Iran than in other countries (Ghandehari, 2016). Over 2.7 million people die from ischemic stroke each year (Lindsay et al., 2019). The mortality rate is affected by factors ranging from admission delay to patient age and comorbidities. Notably, early admission to an intensive care unit and rapid specific diagnosis can significantly improve functional results and reduce mortality (Xian et al., 2011). Survival analyses support clinical prognosis by employing historical data to estimate the mortality risk among patients who suffer a specific illness. Accurate predictions of stroke-caused death risk can help clinicians and hospital administrators take necessary hospital management measures (Smith et al., 2013).

Traditionally, statistical methods such as the Kaplan-Meier test and Cox's proportional hazards model have been employed to predict patient survival. However, novel methods are available today for this purpose. A large number of variables must be analysed simultaneously for prediction of stroke mortality, which limits the use of traditional statistical methods. Data mining methods have recently received particular attention in clinical medicine. In data mining, computer algorithms are employed to extract patterns from large datasets to improve the accuracy and precision of prediction models (Arslan et al., 2016). A data mining algorithm learns from duplicate inputs to discover hidden relationships between data and make predictions on unseen examples (Ni et al., 2018).

A study to predict medical outcomes in patients with sub-arachnoid haemorrhage (SAH) using machine learning techniques in the Waikato Environment for Knowledge Analysis (Weka) was conducted by Paula de Toledo et al. (2009). They analysed two datasets: a set of 441 training data items and another one consisting of 192 data items for external validation. They employed decision tree algorithms such as C4.5, fast decision tree learner, partial decision trees, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), nearest neighbour with generalization, and ripple down rule learner. A model was developed based on the first dataset. The results showed that C4.5 had the best performance based on kappa (0.625) and AUC<sup>1</sup> (84.10). The model was evaluated on the second dataset, and the results indicated that C4.5 outperformed others with kappa and AUC values of 837 and 0.55, respectively. Also, in a work done by Peng et al. (2010) the 30-day mortality rate of patients with SAH was predicted in Weka. The data of 423 patients were analysed by using random forests (RF), artificial neural network (ANN), support vector machine (SVM), and logistic regression (LR). These algorithms were compared to each other in terms of AUC, sensitivity, specificity, accuracy, positive predictive value (PPV) and negative predictive value (NPV). The results showed that RF acquired the highest AUC value (87). In another study by Celik et al. (2014) the 10-day mortality rate in patients with stroke was predicted by using neural networks and multivariate statistical methods. The data of 968 stroke patients were collected and analysed in MATLAB. The results indicated that the highest accuracy was obtained from the multivariate discriminant analysis (MDA) in the haemorrhagic group (87.8%) and from LR in the ischemic group (80.9%). In the study conducted by Ho et al. (2014) to predict post-stroke survival, the imbalanced target variable was balanced by using Synthetic Minority Oversampling Technique (SMOTE). Also, the data mining techniques, including naïve Bayes (NB), SVM, decision trees, RF, and LR in RapidMiner were implemented. Data

---

<sup>1</sup> It means area under the ROC curve.

related to 778 stroke patients were analysed. The results showed that the F1 score (0.33) and c-statistic (0.868) of SVM exhibited the best performance. Also, an independent dataset (39 records) was obtained as an external validator to test the selected classifier. The performance of the selected classifier (SVM) improved after implementation on the independent dataset. Cheon et al. (2019) employed deep learning to predict the mortality rate of patients with stroke. They collected data on 15,099 stroke patients and analysed them by using RF, AdaBoost classifier, Gaussian naïve Bayes, the k-nearest neighbours (KNN), SVM and deep learning in Keras. Sensitivity, specificity, accuracy, AUC and PPV of these algorithms were measured in their study. The AUC results showed that deep learning (83.50%) outperformed other algorithms.

All the above-mentioned studies about prediction of stroke mortality by data mining algorithms have been conducted in high- and upper-middle income countries such as the USA (Ho et al., 2014), Spain (Paula de Toledo et al., 2009), South Korea (Cheon et al., 2019), Taiwan (Peng et al., 2010) and Turkey (Celik et al., 2014). Unfortunately, there is a noticeable lack of studies related to this domain in low- and lower-middle income countries such as Iran.

Considering the rapid increase of stroke incidence in low- and middle-income countries, this study aims to employ the data mining techniques SVM, decision trees, LR, and RIPPER to predict the survival of Iranian stroke patients and to extract corresponding decision rules.

## 2 Materials and methods

The study used a retrospective registry examination design. The endpoint for the analyses was death at hospital discharge. The variables used for prediction were age, sex, smoking history, diabetes mellitus (DM) history, hypertension history, congestive heart disease history, coronary artery problem history, atrial fibrillation (AF) history, cerebrovascular accident (CVA) history, hospital-acquired complications (HACs), cholesterol history, and triglycerides history. Data for the predictive variables were collected at the time of hospital admission. The first stage of the study was carried out on data from the Imam Khomeini Teaching Hospital (dataset A) and the second external validation stage on data from the Imam Reza Hospital (dataset B). The required data were collected retrospectively from medical records of patients with stroke in the period 2011-2017 (see tables 1 and 2).

**Table 1.** General statistics of dataset A variables.

Categorical variables	Distinct values			
Sex	Female		Male	
	Frequency	Percentage	Frequency	Percentage
	2005	48.5%	2133	51.5%
Smoking history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	1035	24.9%	3114	75.1%
Diabetes mellitus (DM) history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	1011	24.4%	3138	75.6%
Hypertension history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	2332	56.2%	1817	43.8%
Congestive heart disease history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	748	18%	3401	82%
Coronary artery problem history	Yes		No	
	Frequency	Percentage	Frequency	Percentage

	231	5.6%	3918	94.4%
Atrial fibrillation (AF) history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	167	4%	3982	96%
Cerebrovascular accident (CVA) history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	1213	29.2%	2936	70.8%
Hospital-acquired complications (HACs) problem	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	707	17%	3442	83%
High cholesterol history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	707	27.3%	1881	72.7%
High triglycerides history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	397	15.3%	2190	84.7%
Body mass index (BMI)	Level1		Level2	Level3
	Frequency	Percentage	Frequency	Percentage
	1002	91.3%	58	5.2%
Length of stay (LOS)	0-14 days		15-60 days	> 61 days
	Frequency	Percentage	Frequency	Percentage
	3498	84.4%	551	13.3%
<b>Numerical variables</b>	Range		Mean	Standard Deviation (SD)
Age	(18-101)		68.19	14.52
Severity of stroke	(0-8)		2.28	1.3

**Table 2.** General statistics of dataset B variables.

Categorical variables	Distinct values			
Gender	Female		Male	
	Frequency	Percentage	Frequency	Percentage
	122	52.1%	112	47.9%
Smoking history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	54	24.4%	167	75.6%
Diabetes mellitus (DM) history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	64	27.4%	170	72.6%
Hypertension history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	133	56.8%	101	43.2%
Congestive heart disease history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	28	12%	206	88%
Coronary artery problem history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	6	2.6%	228	97.4%

Atrial fibrillation (AF) history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	10	4.3%	224	95.7%
Cerebrovascular accident (CVA) history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	64	27.4%	170	72.6%
Hospital-acquired complications (HACs) problem	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	29	12.4%	205	87.6%
High cholesterol history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	34	27%	92	73%
High triglycerides history	Yes		No	
	Frequency	Percentage	Frequency	Percentage
	22	17.6%	103	82.4%
Body mass index (BMI)	Level1		Level2	Level3
	Frequency	Percentage	Frequency	Percentage
	12	7.5%	83	52.2%
Length of stay (LOS)	0-14 days		15-60 days	> 61 days
	Frequency	Percentage	Frequency	Percentage
	209	90.5%	22	9.5%
			0	0
<b>Numerical variables</b>	Range		Mean	Standard Deviation (SD)
Age	(30-93)		69.38	13.16
Severity of stroke	(0-5)		2.12	1

## 2.1 Data collection

Due to the lack of an electronic health record (EHR) system in Iran, the data were collected by reviewing 4149 paper medical records from neurology departments of the Imam Khomeini Teaching Hospital of Urmia (equipped with 635 approved beds) and 234 paper medical records from the Imam Reza Hospital of Urmia (equipped with 256 approved beds). The review of medical records was based on ICD-10 (International Statistical Classification of Diseases and Related Health Problems) codes. Patients aged 0-17 years or those who were not finally diagnosed with either ischemic or haemorrhagic stroke (according to ICD-10 codes) were excluded from the study.

## 2.2 Data structuring

All variables were of the categorical type, except for age and severity of stroke, which were numerical. Cholesterol, triglycerides and LOS were the three variables converted from numerical into categorical using the reduction method. Accordingly, zero and 1 indicated normal and high levels of cholesterol and triglycerides, respectively. In addition, the LOS was classified as three categories of 0-14, 15-60 and over 60 days based on consultation with neurology specialists at the hospital. The severity of stroke and HACs were obtained by merging several variables using the transformation method. The severity of stroke was determined by using the Rapid Arterial occlusion Evaluation (RACE) scale (Lima et al., 2016). The variable HACs was a combination of bedsores, lung infections, urinary tract infection and deep vein thrombosis (DVT).

Since outliers and missing data may affect the final interpretation of knowledge, the number of missing data should be determined so that the necessary measures can be taken to manage them. In this study, some outliers (in 12 records) related to the age and hypertension history variables were detected by the

boxplot and histogram features. The proper values were inserted by a recheck of the paper medical records. These outliers were due to human error when entering values in Excel. The total missing data of dataset A were equal to 5.05%. The missing data were related to the variables of gender (0.26%), high cholesterol history (37.6%), high triglycerides history (37.6%), BMI (73%), LOS (0.04%) and age (0.12%). If more than 50% of records of a variable have missing data, they cannot be effective in analysis and should be excluded from the study. Accordingly, BMI (with 73% missing records) was excluded from this study. The missing data on age, gender and length of stay were substituted by their mean and mode respectively. The missing data on other two variables (high cholesterol history and high triglycerides history) were handled by the multiple imputation method. To perform multiple imputations, the missing data distribution type should be either missing completely at random (MCAR) or missing at random (MAR) (Horton et al., 2007). The missing data distributions for cholesterol and triglycerides were highly related. When data were collected from medical records, it was found out that cholesterol and triglyceride tests had been performed on all patients; however, the test results of some patients had inadvertently not been recorded. Therefore, it was concluded that the missing data distribution of this dataset was of the MAR type. After running the multiple imputation method, five complete datasets were created.

The target variable of dataset A was unbalanced. This means that one value of the target variable has the maximum share of data, whereas other values account for the minority of data. Previous studies have shown that unbalanced learning can lead to predictions with bias and produce incorrect and unreliable results (Gu et al., 2008; Wei et al., 2013). This problem can be solved using methods that balance data such as over-sampling, under-sampling and SMOTE. In over-sampling and under-sampling, the number of minority values of the target variable is increased and reduced, respectively, to balance the target variable. Another technique to balance the target variable is SMOTE. In this technique, new records that are consistent with the records of minority values of the target variable are generated and simulated without changing the records of the majority values to compensate for the imbalance between the values of the target variable (Fernández et al., 2018). The proportion of the target variable values in this study was approximately 5 to 1 (84.6% alive to 15.6% dead), which was balanced by using over-sampling, under-sampling and SMOTE. After balancing the target variable with the mentioned methods, the proportion of its values was 1 to 1 (50% alive to 50% dead). These methods were implemented on the train data and the test data were unchanged.

## 2.3 Data analysis

### 2.3.1 Model construction

The data mining models used in this study were SVM, decision tree, LR, and RIPPER. SVM is a supervised data mining algorithm for classification, prognosis and regression. This method works based on integration of linear algorithms and linear (or nonlinear) nuclear functions (Jeena et al., 2016). In this study, the 10-fold cross-validation technique was employed to conduct SVM.

A decision tree is a simple hierarchical model for managing large volumes of information. The 10-fold cross-validation technique was employed to perform the C5.0 algorithm in this study.

The regression analysis is inherently based on prediction of values of the dependent variable by using several independent variables. As a more general model of linear regression, LR is utilized to predict binary values or variables with multiple discrete values (target variables). In this study, the LR was carried out using the 10-fold cross-validation technique.

RIPPER is a rule-based learner that follows a set of if/then rules. Since these rules are extracted directly from the training datasets, this method can also be called the direct method.

The prediction models (LR, C5 and SVM) were implemented on 20 complete datasets. The target variables of five datasets had not been balanced, whereas the target variables of 15 datasets had been balanced with over-sampling, under-sampling and SMOTE. RIPPER was implemented on five complete datasets.

### 2.3.2 Interpretation and evaluation

The most appropriate models were selected purposively to acquire knowledge. The prediction models used in this study were evaluated in terms of accuracy, sensitivity, specificity, AUC and ROC curves and kappa statistics. For selecting the best representative of each algorithm (LR, C5 and SVM) implemented on 20 datasets, the algorithm that had the best performance in terms of evaluation criteria (equal to or more than 3 criteria) was selected. The criterion to select the best rules using RIPPER from five datasets was the rate of correct classified instances.

### 2.3.3 Data analysis: External validation

In the second stage of the study, dataset A was compared to dataset B for external validation. Dataset A, with 4149 records, and the external validation dataset, with 234 records, were regarded as training and testing datasets, respectively. Then the performance of the best model among three algorithms (LR, C5 and SVM) was evaluated based on accuracy, sensitivity, specificity, AUC and ROC curves and kappa statistics.

## 3 Results

The best performance of LR and C5 resulted from datasets that had been balanced by the over-sampling method. However, the best performance of SVM was due to the under-sampling method. The selected rules extracted by RIPPER were from the first dataset (see Table 3).

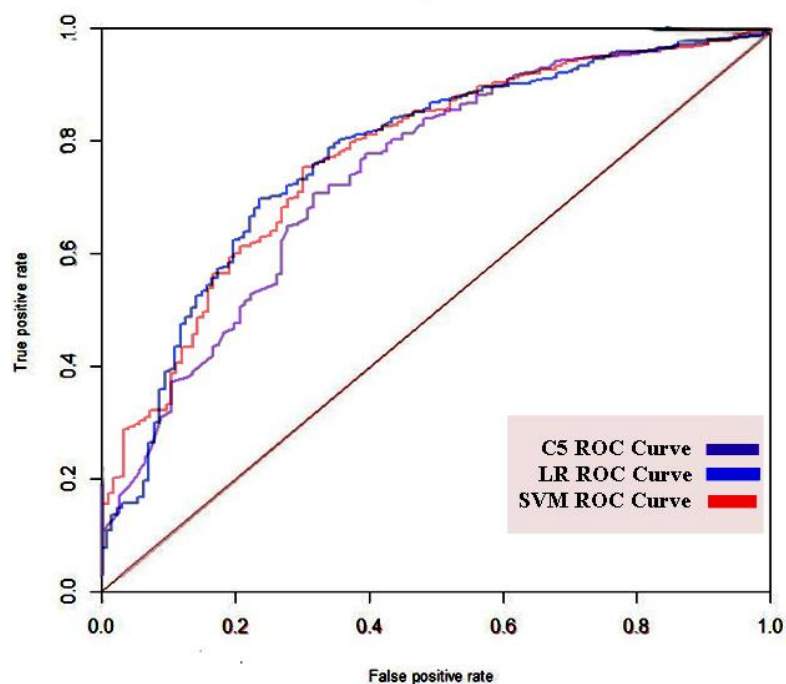
**Table 3.** Performance of algorithms on dataset A.

Performance Algorithm	Accuracy [%]	Sensitivity [%]	Specificity [%]	Kappa [%]	AUC
Decision Tree (C5)	75.39	76.35	70.08	32.87	.75
LR	76.96	79.06	65.35	33.34	.77
Support Vector Machine	75.03	76.21	68.50	31.72	.78

The highest AUC (0.78) was observed for SVM, the highest accuracy, sensitivity and kappa statistic (76.96, 79.06, and 33.34, respectively) were related to LR, and the highest specificity (70.08) belonged to the decision tree. Figure 1 shows the ROC curve of algorithms for predicting the survival of stroke patients in dataset A.

The rules extracted from the RIPPER algorithm for dataset A were:

- A patient aged above 77 years with normal blood pressure, HACs and a stroke severity of 1 is more likely to die.
- A non-smoker patient with HACs and a stroke severity between 1 and 9 is more likely to die.
- A patient aged above 85 years who has been hospitalized for up 60 days is more likely to die.
- A patient aged over 76 years with a stroke severity of 4 or more who has been hospitalized for more than 60 days is more likely to die.
- Otherwise, the patient is more likely to survive.

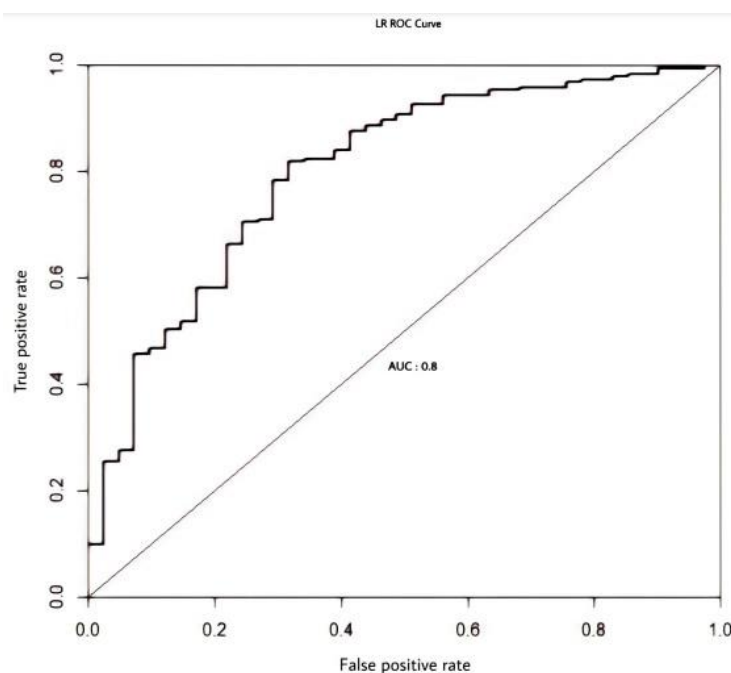


**Figure 1.** ROC curve of predicting algorithms in dataset A.

In the external validation, LR was selected to run on the dataset (B) as the best model of training dataset (A). Table 4 presents evaluation results, and Figure 2 shows the ROC curve of the LR algorithm.

**Table 4.** Performance of best algorithm on dataset B.

Performance Algorithm	Accuracy [%]	Sensitivity [%]	Specificity [%]	Kappa [%]	AUC
LR	79.91	83.94	60.98	39.26	.80



**Figure 2.** ROC curve of LR algorithm in dataset B.



The rules extracted from the RIPPER algorithm for external validation are as follows:

- A patient aged less than 74 years with HACs and a stroke severity of 1 is more likely to die.
- A patient aged over 81 years who has been hospitalized for 15-60 days is more likely to die.

Otherwise, the patient is more likely to survive.

## 4 Discussion

This study aimed to evaluate the performance of data mining techniques (SVM, decision trees and LR) in predicting the survival rate and extracting the rules of stroke patients (using RIPPER) based on clinical data collected from northwestern Iran. When these models were compared in performance on dataset A, it was shown that the LR algorithm outperformed the other algorithms in terms of accuracy (76.96), sensitivity (79.06) and kappa (33.34). Its performance on the external validation dataset improved in all of the evaluation criteria except specificity.

A few studies have been conducted on the prediction of survival in stroke patients using data mining techniques. The accuracy index in some previous studies was higher than that of this study. For instance, it was 87.8 in Celik et al. (2014), 84.03 in Cheon et al. (2019), and 78.50 in Peng et al. (2010). However, it can be stated that the accuracy of algorithms is usually high in datasets with unbalanced values of the target variable due to predicting values of very high frequency. The kappa statistic excludes chance from calculations and provides an accuracy level that cannot be obtained from this simple strategy. However, none of the above studies reported using this indicator.

The AUC values, which establish a balance between sensitivity and specificity, reported by Cheon et al. (2019), Peng et al. (2010) and Paula de Toledo et al. (2009), were higher than the value obtained in this study. This is probably because no balancing algorithm was employed in those studies.

Ho et al. (2014) and Paula de Toledo et al. (2009) also employed external validation in their studies on datasets of 39 and 192 records, respectively. The dataset size of both studies was smaller than that of this study. However, like our work, the performance of the selected algorithm of both studies on the external validation dataset was slightly better.

Following are some of the major differences between this study and previous ones. Some methods were used in this study for missing value management to prevent the loss of a variable or record. None of the previous studies employed such methods. Ho et al. (2014) used SMOTE for balancing, whereas SMOTE, oversampling and under-sampling algorithms were utilized in this study for this purpose. Apart from the study conducted by Cheon et al. (2019) on 15,099 records, the number of records in this study was higher than that of all other previous studies. It is noteworthy that Cheon et al. (2019) collected data from an electronic health records system, whereas paper medical records were reviewed for data collection. MRI and CT scans of patients were also used for data collection in other studies (Paula de Toledo et al., 2009; Peng et al., 2010; Cheon et al., 2019), whereas such documents were not available in medical records of patients to be used in this study. All studies except Celik et al. (2014) and Peng et al. (2010) employed the 10-fold cross-validation technique, which reduces computation time as well as prediction bias. In this study, R was used for running the model. R is an open-source programming language consisting of several libraries in the field of artificial intelligence and data sequence. Other studies (Paula de Toledo et al., 2009; Peng et al., 2010; Celik et al., 2014; Ho et al., 2014; Cheon et al., 2019) employed RapidMiner, MATLAB, Weka and Keras for this purpose.

There were also some similarities between this study and previous ones (Paula de Toledo et al., 2009; Peng et al., 2010; Celik et al., 2014; Ho et al., 2014; Cheon et al., 2019). All of the research variables were similar to those of the previous studies. The indicators used in this study for the evaluation of algorithms (e.g.,

accuracy, sensitivity, specificity, etc.) were the same as those employed in the previous studies. Finally, both this study and the previous ones employed some common algorithms.

The survival of stroke patients has been predicted to be 6 hours (Lewis et al., 2008), 5 days (Counsel et al., 2003), 30 days (Counsell et al., 2002), 3 months (Johnston et al., 2000; Konig et al., 2008), 100 days (Weimar et al., 2002), 6 months (Counsel et al., 2003) and 1 year (Hankey et al., 2000; Li et al., 2012). The traditional statistical methods employed in those studies are different from the methods used in this study.

It can be stated that age, HACs, male gender and severity of disease were the most important factors affecting the survival of stroke patients (Feigin et al., 2016; Wijaya et al., 2019). Moreover, diseases such as diabetes, hypertension, hyperlipidaemia and high cholesterol did not affect the survival of stroke patients (Nam et al., 2012; Salman et al., 2012).

#### 4.1 Limitations and future direction

The first and most important limitation of this study was the use of the RACE index to determine the severity of stroke. The reason for using this index instead of the National Institutes of Health Stroke Scale (NIHSS) index was the lack of access to this index before 2015. If using the index, it should have been removed due to the large amount of missing data. At the same time, the RACE index was used due to its effectiveness. Also the reason for excluding some items such as family history, non-exploitation of brain images (CT scan and MRI) and genetic tests in this study was the incompleteness of medical records as well as the unavailability of brain images in clinical records.

Future studies are recommended to analyse the effects of new variables such as insurance coverage, medications, quantity and quality of hospital services and comorbidities on the survival of stroke patients, include brain images (MRI and CT scans) in their study, and to review more records.

### 5 Conclusion

On a dataset of 4149 records, some data mining-based models were developed to automatically predict the survival of stroke patients concerning clinical data and history of comorbidities. Taking care of stroke patients is a difficult and complicated process that requires emergency clinical decisions in the first stage of the disease. The research findings allow early diagnosis of patients facing the risk of death from a stroke who need additional examinations and appropriate treatment before the disease worsens. The identification of such patients can be helpful in clinical, managerial and hospital decisions. The classification algorithm was employed due to the nature of the variable (survival), which contains two modes of survival and non-survival. The LR algorithm performed better than other ones.

To further ensure the performance of the algorithms, a dataset from a separate hospital was used for external validation. Dataset A (Imam Khomeini Teaching Hospital) and dataset B (Imam Reza Hospital) were regarded as training and testing datasets, respectively. The results of external validation LR outperformed other algorithms with an AUC value of 0.80. The rules extracted by the RIPPER were clinically meaningful and clarify the hidden patterns of the datasets.

### Additional Information and Declarations

**Funding:** This work was funded by the deputy of research, Urmia University of Medical Sciences, Grant No. 96-09-52-3170.

**Conflict of Interests:** The authors declare no conflict of interest.

**Author Contributions:** Z.H.: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology and Visualization. H.L.A.: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision,

Validation and Writing – review & editing. S.N.: Data curation, Investigation, Visualization and Writing – review & editing. B.H.: Conceptualization, Formal analysis, Methodology, Validation and Writing – original draft. T.T.: Formal analysis, Methodology, Validation and Writing – original draft.


**Data Availability:** The data that support the findings of this study are available from the corresponding author.

## References

- Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. *Computer Methods and Programs in Biomedicine*, 130, 87–92. <https://doi.org/10.1016/j.cmpb.2016.03.022>
- Azarpazhooh, M. R., Etemadi, M. M., Donnan, G. A., Mokhber, N., Majdi, M. R., Ghayour-Mobarhan, M., Ghandehary, K., Farzadfard, M. T., Kiani, R., Panahandeh, M., & Thrift, A. G. (2010). Excessive Incidence of Stroke in Iran. *Stroke*, 41(1). <https://doi.org/10.1161/strokeaha.109.559708>
- Boehme, A. K., Esenwa, C., & Elkind, M. S. V. (2017). Stroke Risk Factors, Genetics, and Prevention. *Circulation Research*, 120(3), 472–495. <https://doi.org/10.1161/CIRCRESAHA.116.308398>
- Çelik, G., Baykan, Ö. K., Kara, Y., & Tireli, H. (2014). Predicting 10-day Mortality in Patients with Strokes Using Neural Networks and Multivariate Statistical Methods. *Journal of Stroke and Cerebrovascular Diseases*, 23(6), 1506–1512. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2013.12.018>
- Cheon, S., Kim, J., & Lim, J. (2019). The Use of Deep Learning to Predict Stroke Patient Mortality. *International Journal of Environmental Research and Public Health*, 16(11). <https://doi.org/10.3390/ijerph16111876>
- Counsell, C., Dennis, M., McDowall, M., & Warlow, C. (2002). Predicting Outcome After Acute and Subacute Stroke. *Stroke*, 33(4), 1041–1047. <https://doi.org/10.1161/hs0402.105909>
- Counsel, C., Dennis, M. S., Lewis, S., Warlow, C., FOOD Trial Collaboration. Feed Or Ordinary Diet. (2003). Performance of a Statistical Model to Predict Stroke Outcome in the Context of a Large, Simple, Randomized, Controlled Trial of Feeding. *Stroke*, 34(1), 127–133. <https://doi.org/10.1161/01.str.0000044165.41303.50>
- de Toledo, P., Rios, P. M., Ledezma, A., Sanchis, A., Alen, J. F., & Lagares, A. (2009). Predicting the Outcome of Patients With Subarachnoid Hemorrhage Using Machine Learning Techniques. *IEEE Transactions on Information Technology in Biomedicine*, 13(5), 794–801. <https://doi.org/10.1109/titb.2009.2020434>
- Feigin, V. L., Roth, G. A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., Mensah, G. A., Norrving, B., Shiue, I., Ng, M., Estep, K., Cercy, K., Murray, C. J. L., & Forouzanfar, M. H. (2016). Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet Neurology*, 15(9), 913–924. [https://doi.org/10.1016/s1474-4422\(16\)30073-4](https://doi.org/10.1016/s1474-4422(16)30073-4)
- Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905. <https://doi.org/10.1613/jair.1.11192>
- Gebreyohannes, E. A., Bhagavathula, A. S., Abebe, T. B., Seid, M. A., & Haile, K. T. (2019). In-Hospital Mortality among Ischemic Stroke Patients in Gondar University Hospital: A Retrospective Cohort Study. *Stroke Research and Treatment*, 2019, 1–7. <https://doi.org/10.1155/2019/7275063>
- Ghandehari, K. (2016). Epidemiology of Stroke in Iran. *Galen Medical Journal*, 5(supplement 1), 3–9.
- Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data Mining on Imbalanced Data Sets. In *2008 International Conference on Advanced Computer Theory and Engineering* (pp. 1020–1024). IEEE. <https://doi.org/10.1109/ICACTE.2008.26>
- Hankey, G. J., Jamrozik, K., Broadhurst, R. J., Forbes, S., Burvill, P. W., Anderson, C. S., & Stewart-Wynne, E. G. (2000). Five-Year Survival After First-Ever Stroke and Related Prognostic Factors in the Perth Community Stroke Study. *Stroke*, 31(9), 2080–2086. <https://doi.org/10.1161/01.str.31.9.2080>
- Ho, K.C., Speier, W., El-Saden, S., Liebeskind, D.S., Saver, J.L., Bui, A.A., Arnold, C.W. (2014). Predicting discharge mortality after acute ischemic stroke using balanced data. In *AMIA Annual Symposium proceedings, 2014* (pp. 1787–1796). AMIA.
- Horton, N. J., & Kleinman, K. P. (2007). Much Ado About Nothing. *The American Statistician*, 61(1), 79–90. <https://doi.org/10.1198/000313007x172556>
- Hosseini, A. A., Sobhani-Rad, D., Ghandehari, K., & Benamer, H. T. (2010). Frequency and clinical patterns of stroke in Iran – Systematic and critical review. *BMC Neurology*, 10(1). <https://doi.org/10.1186/1471-2377-10-72>
- Jeena, R. S., & Kumar, S. (2016). Stroke prediction using SVM. In *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies* (pp. 600–602). IEEE. <https://doi.org/10.1109/ICCCCT.2016.7988020>
- Johnston, K. C., Connors, A. F., Wagner, D. P., Knaus, W. A., Wang, X.-Q., & Haley, E. C. (2000). A Predictive Risk Model for Outcomes of Ischemic Stroke. *Stroke*, 31(2), 448–455. <https://doi.org/10.1161/01.str.31.2.448>
- Köhlgl, R., Ziegler, A., Bluhmki, E., Hacke, W., Bath, P. M. W., Sacco, R. L., Diener, H. C., & Weimar, C. (2008). Predicting Long-Term Outcome After Acute Ischemic Stroke. *Stroke*, 39(6), 1821–1826. <https://doi.org/10.1161/strokeaha.107.505867>

- Lewis, S. C., Sandercock, P. A., & Dennis, M. S.** (2008). Predicting outcome in hyper-acute stroke: validation of a prognostic model in the Third International Stroke Trial (IST3). *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4), 397–400. <https://doi.org/10.1136/jnnp.2007.126045>
- Li, W.-J., Gao, Z.-Y., He, Y., Liu, G.-Z., & Gao, X.-G.** (2011). Application and Performance of Two Stroke Outcome Prediction Models in a Chinese Population. *PM&R*, 4(2), 123–128. <https://doi.org/10.1016/j.pmrj.2011.08.669>
- Lima, F. O., Silva, G. S., Furie, K. L., Frankel, M. R., Lev, M. H., Camargo, É. C. S., Haussen, D. C., Singhal, A. B., Koroshetz, W. J., Smith, W. S., & Nogueira, R. G.** (2016). Field Assessment Stroke Triage for Emergency Destination. *Stroke*, 47(8), 1997–2002. <https://doi.org/10.1161/strokeaha.116.013301>
- Lindsay, PM., Norrving, B., Sacco, R.L., Brainin, M., Hacke, W., Martins, S.H., Pandian, J., & Feigin, V.** (2019). *World Stroke Organization (WSO): Global Stroke Fact Sheet 2019*. Retrieved November 15, 2021, from <https://www.world-stroke.org>
- Nam, H. S., Kim, H. C., Kim, Y. D., Lee, H. S., Kim, J., Lee, D. H., & Heo, J. H.** (2012). Long-Term Mortality in Patients With Stroke of Undetermined Etiology. *Stroke*, 43(11), 2948–2956. <https://doi.org/10.1161/strokeaha.112.661074>
- Ni, Y., Alwell, K., Moomaw, C. J., Woo, D., Adeoye, O., Flaherty, M. L., Ferioli, S., Mackey, J., De Los Rios La Rosa, F., Martini, S., Khatri, P., Kleindorfer, D., & Kissela, B. M.** (2018). Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. *PLOS ONE*, 13(2), e0192586. <https://doi.org/10.1371/journal.pone.0192586>
- Peng, S.-Y. ., Chuang, Y.-C. ., Kang, T.-W. ., & Tseng, K.-H.** (2010). Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. *European Journal of Neurology*, 17(7), 945–950. <https://doi.org/10.1111/j.1468-1331.2010.02955.x>
- Salman, R. R., Delbari, A., & Tabatabae, S. S.** (2012). Stroke rehabilitation: principles, advances, early experiences, and realities in Iran. *Journal of Sabzevar University of Medical Sciences*, 19(2), 96–108.
- Smith, E. E., Shobha, N., Dai, D., Olson, D. M., Reeves, M. J., Saver, J. L., Hernandez, A. F., Peterson, E. D., Fonarow, G. C., & Schwamm, L. H.** (2013). A Risk Score for In-Hospital Death in Patients Admitted With Ischemic or Hemorrhagic Stroke. *Journal of the American Heart Association*, 2(1). <https://doi.org/10.1161/jaha.112.005207>
- Wei, Q., & Dunbrack, R. L.** (2013). The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, 8(7), e67863. <https://doi.org/10.1371/journal.pone.0067863>
- Weimar, C., Ziegler, A., König, I. R., & Diener, H.-C.** (2002). Predicting functional outcome and survival after acute ischemic stroke. *Journal of Neurology*, 249(7), 888–895. <https://doi.org/10.1007/s00415-002-0755-8>
- Wijaya, HR., Supriyanto, E., Salim, MIM., Siregar, KN., & Eryando, T.** (2019). Stroke management cost: Review in Indonesia, Malaysia and Singapore. *AIP conference proceedings*, 2092(1), 030022.
- Xian, Y., Holloway, R.G., Chan, P.S., Noyes, K., Shah, M.N., Ting, H.H., Chappel, A.R., Peterson, E.D., & Friedman, B.** (2011). Association Between Stroke Center Hospitalization for Acute Ischemic Stroke and Mortality. *The Journal of the American Medical Association*, 305(4), 373–380. <https://doi.org/10.1001/jama.2011.22>

---

**Editorial record:** The article has been peer-reviewed. First submission received on 20 September 2021. Revisions received on 31 October 2021, and 17 November 2021. Accepted for publication on 18 November 2021. The editor in charge of coordinating the peer-review of this manuscript and approving it for publication was Zdenek Smutny .

---

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951

---