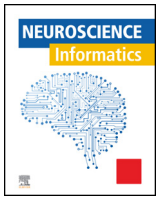
Neuroscience Informatics 5 (2025) 100181

Contents lists available at ScienceDirect

Neuroscience Informatics

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



Original article

Understanding risk factors of post-stroke mortality ☆

David Castro [a], Nuno Antonio [a,*], Ana Marreiros [b], Hipólito Nzwalo [b]

[a] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal
[b] Faculty of Medicine and Biomedical Sciences, Universidade of Algarve, Portugal

ARTICLE INFO

Article history:
Received 27 October 2024
Received in revised form 23 November 2024
Accepted 25 November 2024

Keywords:
Risk factors analysis
Stroke
Mortality
Machine learning
Modified Rankin scale

ABSTRACT

Stroke is one of the leading causes of death worldwide. Understanding the risk factors for post-stroke mortality is crucial for improving patient outcomes. This study analyzes and predicts post-stroke mortality using the modified Rankin Scale (mRS), a functional neurological evaluation scale. Several Machine Learning models were developed and assessed using a dataset of 332 stroke patients from Hospital de Faro, Portugal, from 2016 to 2018. The Random Forest model outperformed others, achieving an accuracy of 98.5% and a recall of 91.3. Twenty-four risk factors were identified, with stroke severity as the most critical. These findings provide healthcare professionals with valuable tools for early identification and intervention for high-risk stroke patients, enabling informed decision-making and customized treatment plans. This research advances healthcare predictive analytics, offering a precise mortality prediction model and a comprehensive analysis of risk factors, potentially improving clinical outcomes and reducing mortality rates. Future applications could extend to patient monitoring and management across various medical conditions.

1. Introduction

Stroke is a neurological deficit caused by an acute focal injury of the central nervous system due to vascular reasons. It is the second leading cause of death worldwide and the primary cause of disability in adults [8,16,33,34,37,67,71].

In Portugal, despite a recent decline in stroke mortality rates, cardiovascular diseases persist as the primary cause of death [41,52]. In 2011, Portugal displayed a standardized mortality rate of 62.64 per 100,000 inhabitants attributable to cerebral vascular diseases, which ranked Portugal as the third European country with the highest mortality rate in this category. The impact of strokes resulted in 13,020 deaths and 14,379 potential years of life lost [41]. In 2019, according to the Global Burden of Diseases, Injuries, and Risk Factors study, stroke accounted for 16,695 deaths, representing 14.34% of all deaths attributed to diseases and injuries in the country that year [26].

The risk of death between four weeks and one year post a first stroke is five times higher in patients with non-fatal stroke than in the general population, with a subsequent twofold increase after one year [16]. Understanding the risk factors of post-stroke mortality is essential for personalized and early effective patient care, prevention strategies, resource allocation, advancements in stroke research, and healthcare policies and prognostication [17,32,35,56,71].

Integrating Machine Learning (ML) in biomedical research, especially stroke medicine, is gaining popularity for its advanced analysis and predictive capabilities [60]. By surpassing traditional regression methods, ML algorithms excel in predicting post-stroke mortality risk factors [56]. Besides patterns' identification in complex datasets and human bias reduction, these techniques enable efficient analysis of extensive datasets, guiding personalized and precision medicine in stroke care [16]. ML models offer enhanced performance in prognostic modeling by accommodating more predictors, utilizing an agnostic approach, and handling multi-dimensional correlations [56]. Moreover, the collaborative use of classification algorithms and association rule mining, as seen in [4] this example, helps better understand how a predictor can influence the outcome.

The modified Rankin Scale (mRS) is the most widely used functional outcome measure in stroke clinical trials [21,73,77]. It evaluates the degree of disability in stroke patients on an ordinal scale, with seven ordered but unequally spaced categories, ranging from a score of 0 (no symptoms) to 6 (death) [15,37,66]. In this sense, this study aims to understand the risk factors for post-stroke mortality, providing possible explanations through ML techniques. As such, the development of an optimal ML algorithm (objective 1)

☆ This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.
* Corresponding author at: NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312, Lisbon, Portugal.
E-mail address: nantonio@novaims.unl.pt (N. Antonio).

https://doi.org/10.1016/j.neuri.2024.100181

and the identification and analysis of the risk factors (objective 2) are addressed by the prediction of the mRS. The identification of risk factors has the potential to allow clinical teams to improve the outcomes, thus reducing mortality.

This study uses a dataset of 332 stroke patients treated at Centro Hospitalar Universitário do Algarve - Hospital de Faro, Portugal, from 2016 to 2018, encompassing 80 distinct variables. The study involves comprehensive data analysis, pre-processing, feature selection, and testing of various ML models, including tree-based models, artificial neural networks (ANN), probabilistic models, kernel machines, and ensemble methods. Besides supervised learning, other techniques are deployed to reveal strong relationships between attributes, identify main features, reduce dimensionality, populate minority classes, and provide output explanations, ultimately leading to a mRS score of 6. From a healthcare perspective, mortality prediction is a valuable tool aiding clinicians in prognosis, care planning, therapy selection, rehabilitation coordination, counseling, hospital outcome comparisons, and performance assessment related to stroke mortality [16].

2. Literature review

Much has been written about the prognosis after stroke, including many studies on short-term survival, recurrence, recovery, life expectancy, and mortality [33,57]. In this chapter, we address the post-stroke mortality risk factor identification-related work and other risk factors related to stroke. Chronologically, we discuss the identified factors and the different models applied by synthesizing the related work.

The literature review uses the Scopus database as the primary resource, ordered by relevance. Searches are focused on key terms such as 'Stroke,' 'Stroke Mortality,' 'Risk Factors of Post-Stroke Mortality,' 'Modified Rankin Scale,' and 'Predicting the Modified Rankin Scale' to identify relevant studies. Additionally, to refine the focus on the Portuguese population, the search queries are supplemented with the inclusion of the term 'Portugal.' Upon initial identification of Portuguese-specific literature, further investigation is undertaken by scrutinizing references from the most relevant articles.

2.1. Post-stroke mortality risk factors identification

The exploration of post-stroke mortality spans several decades and employs diverse methodologies, shedding light on critical risk factors associated with poor outcomes. Table 1 summarizes the potential factors leading to death after stroke episodes selected in some essential research contributions over the years and the techniques used to identify them.

Among the twenty-four studies examined, older age and stroke severity, such as decreased level of consciousness and dependency, have emerged as the predominant predictors of mortality following a stroke, being cited in 18 instances. Subsequently, atrial fibrillation (AF), diabetes, and prior stroke were featured as predictors in 10 instances. Gender is mentioned nine times, pneumonia eight times, smoking status and stroke type are identified seven times each, and hypertension is associated with mortality in 6 instances.

In methodological terms, the studies use various statistical approaches, from the traditional Cox and Logistic Regression (LR) models to the advanced ML techniques used more recently. The evolution of methodologies over time underlines the ongoing efforts to improve prediction accuracy and understand the multifactorial nature of post-stroke mortality.

Five articles related to death after stroke are analyzed [2,24,46,47,63] about the Portuguese population. Of these, most focused on the analysis of hemorrhagic stroke, leaving a research gap for the analysis of both types of strokes (hemorrhagic and ischemic) in Portugal. There is, therefore, an urgent need to explore more factors and give consistency to the conclusions reached so far. Furthermore, although the articles are recent, they employ traditional prediction models, such as the Cox model, and do not follow the evolution towards ML models, as we observed in most recent literature.

Although most scientific papers successfully identify singular risk factors for long - and short-term mortality following a stroke using significance tests, they often fail to explore the combined effects of these factors, as they solely concentrate on their independent contributions. They also tend to concentrate on identifying risk factors without delving into how each risk factor specifically impacts mortality after a stroke. As a result, a research gap exists in the analysis and comparison of the effects of individual risk factors on mortality.

2.2. Stroke-related risk factors identification

The relationship between stroke risk factors and post-stroke mortality is intricate and interconnected. Various risk factors contributing to the likelihood of experiencing a stroke can be the same risk factors leading to post-stroke mortality [2]. Following stroke, these risk factors amplify the complications inherent to the event, such as infections and secondary strokes [75]. The coexistence of multiple risk factors often creates a cumulative impact, heightening the overall risk of mortality post-stroke. Moreover, shared physiological pathways and interdependencies among these risk factors further complicate the picture, emphasizing the importance of simultaneously addressing a spectrum of factors for effective stroke prevention and improved post-stroke outcomes [2].

Several scientific papers have delved into the complex landscape of stroke, exploring risk factors, consequences, and predictive models. Table 2 summarizes the potential factors related to stroke episodes selected in some vital research contributions and the techniques used to identify them over the years.

The examination of stroke-related factors from diverse studies reveals nuanced insights into this complex condition, with contributors such as hypertension, older age, diabetes, AF, smoking status, gender, and stroke severity significantly contributing to stroke risk. Unique risk factors for specific populations are also identified [74]. In parallel, predictive methodologies utilized diverse models, ranging from LR to ML algorithms, emphasizing the need for tailored approaches [17,41]. These findings collectively underscore the multifaceted nature of stroke, necessitating a comprehensive understanding of various risk factors and diverse analytical approaches for accurate prediction and management.

Considering both tables, we can state that other stroke-related consequences share most of the predictors of post-stroke mortality, meaning that both subjects should be considered together in the analysis. These common factors are older age, AF, diabetes, gender, smoking status, and hypertension.

On the one hand, by analyzing Table 2, we are adding confidence to the predictors and extending the considered literature on models and risk factors to serve as a base for our following sections. For example, we now know other ML techniques used in predicting stroke, such as Shapley additive explanations (SHAP), to interpret the predictors that still were not used in predicting post-stroke mortality factors in section 2.1 [17,36]. On the other hand, we are highlighting the possibility of having predictors of post-stroke mortality that may not be predictors of stroke episode or recurrence, such as stroke-associated pneumonia (SAP), stroke severity, cancer, and kidney disease [16,30,45].

Table 1
Related work on techniques and risk factors for post-stroke mortality.

Reference	Factors	Methods
[9]	Stroke severity, age, marital status, stroke history, and ethnic group.	Cox model
[29]	Stroke severity, previous stroke, and age.	Logistic Regression (LR)
[31]	Rankin score, history of myocardial infarction, stroke type, diabetes, smoking, meat, alcohol, and aspirin.	Cox model
[6]	Gender, age, stroke severity, stroke type, ischemic heart disease, hypertension, Atrial Fibrillation (AF), diabetes, previous stroke, smoking, and alcohol.	Cox model
[19]	Not using beta-blockers, age, stroke severity, fasting glucose, total cholesterol level, ischemic heart disease, AF, smoking, total anterior circulation infarct, and Stroke-associated pneumonia (SAP).	Kaplan–Meier statistics, Cox model, LR
[22]	High mRS, gender, age, diabetes, smoking, hypertension therapy, AF, and depressed mood.	Cox model
[61]	Age, and AF.	Cox model
[45]	Acute ischemic stroke, SAP, intracerebral hemorrhage, recurrent stroke, myocardial infarction, cancer, age, hypertension, coronary disease, NIHSS, undetermined stroke etiology, comorbidities, hyperglycemia, AF, signs of ischemia, dense artery sign, proximal vessel occlusion, and thrombolysis.	Univariate and Multivariate comparisons
[11]	Age, gender, hypertension, diabetes, smoking, blood pressure, stroke severity, stroke type, AF, glycaemia, heart failure, coronary heart disease, myocardial infarction, history of cerebrovascular disease, transient ischemic attack, and hemorrhagic characteristics.	LR, ANN, multivariate discriminant analysis
[20]	Short-term risk: medication, hypocholesterolemia and previous stroke. Intermediate-term risk: age, AF and heart failure, socio-demographics, stroke type, and severity.	Naïve Bayes, Decision Tree (DT), LR
[39]	SAP, and urinary infections.	Poisson and Cox models
[47]	Age, vitamin K antagonists, Glasgow Coma Scale (GCS), hematoma volume, intraventricular dissection, and pneumonia.	Kaplan–Meier and LR
[46]	Age, stroke severity, discharged to nursing units, re-hospitalization, pneumonia, cardiovascular complications, non-respiratory infections; neoplasia; and recurrent stroke.	Kaplan–Meier and Cox model
[23]	Dysphagia, and SAP.	Cox models
[30]	AF, age, myocardial infarction, gender, chronic heart failure, smoking, cancer, previous stroke, time from onset to admission, medication history, NIHSS, statin, antiplatelet, anticoagulation, systolic/diastolic blood pressure, previous diseases, hemoglobin, pre-stroke mRS, white blood cell count, hypertension, platelet count, diabetes, prothrombin time, hypercholesterolemia, glycaemia, and metabolic syndrome.	ANN, Random Forest, LR
[57]	Age, gender, and mRS.	-
[2]	Gender, age, functional status, and internal hospitalization length of stay.	Kaplan–Meyer method, Cox model
[56]	Age, length of the follow-up, stroke severity, time from onset to rehabilitation, renal dysfunction, AF, and diabetes.	Random Forest, SMOTE
[63]	Age, gender, stroke severity, and pneumonia.	Kaplan–Meier, LR
[64]	AF.	Cox model
[24]	Unconsciousness, time from onset to hospital admission, hematoma volume, intraventricular extension, and emergency time.	LR
[76]	SAP.	Bayesian
[69,70]	Age, gender, symptom onset time, hypertension, diabetes, smoking, alcohol, antiplatelet, anticoagulation, GCS, and systolic/diastolic blood pressure.	Support Vector Machine (SVM), LR
[16]	Age, gender, severe stroke subtype, glucose, AF, coronary artery disease, congestive heart failure, cancer, dementia, kidney disease, and dependency pre-stroke.	ANN, SVM, Random Forest, SMOTE

LR – Logistic Regression; ANN – Artificial Neural Networks; DT – Decision Tree; SMOTE – Synthetic Minority Oversampling Technique; SVM – Support Vector Machine; GCS - Glasgow Coma Scale; AF – Atrial Fibrillation; SAP – Stroke-associated pneumonia.

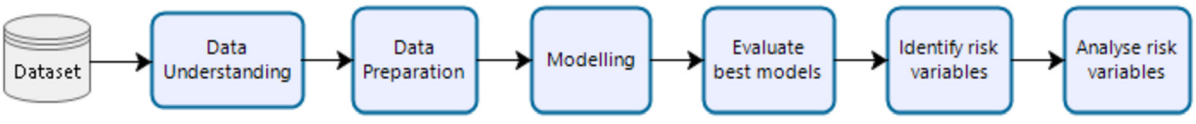


Fig. 1. Proposed protocol.

3. Methodology

This section presents the general framework for this study. After focusing on the business understanding phase, we are now focusing on the following three phases of the Cross Industry Standard Process for Data Mining: Data Understanding, Data Preparation, and Modelling [12]. Fig. 1 provides the design diagram for this study.

3.1. Data understanding

The dataset for this study consists of 332 stroke patients admitted to Centro Hospitalar Universitário do Algarve—Hospital de Faro, Portugal, from 2016 to 2018. It covers 80 different variables, including automatic administrative data, sociodemographics, usual medication, pre-stroke history, characteristics on admission, complications during hospitalization, and biomarkers on admission.

Table 2
Related work on techniques and risk factors connected to stroke.

Reference	Prediction	Factors	Methods
[52]	Length of hospital stay	Gender, paralysis, and intracerebral hemorrhage.	LR
[49]	Stroke	Hypertension, diet risk score, regular physical activity, smoking, diabetes, alcohol, waist-to-hip ratio, psychosocial stress, depression, and cardiac causes.	LR
[10]	SAP	Age, gender, neurologic deficit severity, and longer hospitalization.	-
[75]	Recurrent ischemic stroke	History of coronary heart disease, severe stenosis or occlusion of large cerebral artery, and multiple acute cerebral infarcts.	Cox model
[41]	Stroke	Age, gender, race, family history, Hypertension, AF, dyslipidemia, smoking habits, and diabetes.	-
[13]	Stroke	Genetic risk factors: AF, coronary artery disease, blood pressure regulation, pericyte and smooth muscle cell development, coagulation, carotid plaque formation, and neuro-inflammation.	-
[48]	Stroke	Hypertension of 140/90 mm Hg or higher, regular physical activity, apolipoprotein, diet, waist-to-hip ratio, psychosocial factors, current smoking, cardiac causes, alcohol consumption, and diabetes.	LR
[3]	Stroke in young adults	Hypertension, low physical activity, smoking, and alcohol consumption.	-
[28]	Stroke	Diabetes, hypertension, age, and gender.	SVM, LR, Tree, and Ensemble
[78]	Recurrent ischemic stroke	Coronary heart disease, AF, hyperlipidemia, hyperhomocysteinemia, improper diet, overwork, emotional excitement, body mass index, retinal vessel structures and arterial-venous ratio.	LR
[35]	Stroke in farmers	Gender, age, personal history of hypertension, diabetes, current smoking, high γ-glutamyl transferase, and metabolic syndrome components.	LR
[58]	Stroke	Diabetes, age, duration of type 2 diabetes, estimated glomerular filtration rate, blood pressure, lipid levels, body mass index, uric acid, and glycosylated hemoglobin A1c.	LR, LASSO
[59]	Stroke in dialysis patients	Diabetic nephropathy, hypertension, diabetes, and dyslipidemia, dialysis introduction, elderly, past stroke history, AF, hyperparathyroidism, hyperhomocysteinemia, obesity, serum albumin value, serum phosphate levels, anemia, hypocalcemia, hypercytokinemia, high-sensitivity C-reactive protein, concentration of asymmetric dimethylarginine, disordered lipid metabolism, dialysis technology, vascular access, use of anti- coagulant.	-
[74]	Stroke recurrence	Age, arterial hypertension, AF, diabetes, hyperlipidaemia, past transient ischemic attacks, cerebral atherosclerosis, white matter lesions, and cardiac disease and retinal characteristics.	LR
[27]	Infection post-stroke	Dysphagia, vitamin D deficiency, age, cigarette smoking, and the use of proton pump inhibitors, H2 receptor antagonists, and benzodiazepines.	-
[67]	SAP	Cerebral hemorrhage, indwelling nasogastric tube, and high neutrophil-to-lymphocyte ratio.	-
[17]	Stroke	Age, heart disease, average glucose level, and hypertension.	PCA, SMOTE, ANN, DT, RF, SVM, LASSO and ElasticNet
[34]	Stroke	Carotid intima-media thickness, age, gender, hypertension, cigarette smoking, diabetes, and hypercholesterolemia.	-
[40]	Stroke in elderly woman	AF, hormone therapy, psychosocial risk factors, and cognitive impairment.	-
[76]	SAP	Mechanical ventilation, AF, pre-existing respiratory disease, smoking, pre-existing heart disease, stroke severity, stroke-induced immunodepression and dysphasia.	-
[62]	SAP	GCS, prolonged emergency room stay and hyperactive delirium.	LR
[71]	length of stay in hospital	NIHSS, AF, receiving thrombolytic therapy, history of hypertension, diabetes, previous stroke history.	ANN
[69]	Stroke	Total cholesterol, serum creatinine, systolic blood pressure, age, heart disease, white blood cells, hypertension and history of stroke.	Bagging and Boosting
[36]	Stroke	Hypertension, history of transient ischemia, and history of stroke.	LR, SVM, Light Grading Boosting, XGBoost, SHAP

LR – Logistic Regression; SVM – Support Vector Machine; LASSO – Least Absolute Shrinkage and Selection Operator; PCA – Principal Component Analysis; SMOTE – Synthetic Minority Oversampling Technique; ANN – Artificial Neural Networks; DT – Decision Tree; RF – Random Forest; XGBoost – Extreme Gradient Boosting; SHAP - SHapley Additive exPlanations; GCS - Glasgow Coma Scale; AF – Atrial Fibrillation; SAP – Stroke-associated pneumonia; NIHSS – National Institutes of Health Stroke Scale.

Table 3 presents the description of each variable. All patient data has been desensitized.

Fig. 2 shows the skewed class proportions that make this dataset imbalanced. For example, according to the mRS, nine times more people are evaluated at level 4 than at level 1. Nevertheless, the most important category for this study is well represented, with more than a seventh of the data. More specifically, of the 332 patients, 61 died a few days after having a stroke.

3.2. Data preparation

The data preparation phase has five core tasks: selecting, cleaning, constructing, integrating, and formatting data [55]. Fig. 3 summarizes the procedures used to prepare the data for the modeling part.

The dataset was initially cleaned by removing only columns with missing values. Subsequently, the last two columns containing comments were merged. The “ID_patient” column was dropped, and the “Episódio” column was set as the index. Columns and their respective values were translated into English, with “N/D” and “N/A” replaced with appropriate missing values.

New features have been created from columns with dates because these alone do not help to predict mortality, but perhaps, as time intervals, they can lead us to some conclusions. These are “DU_HAD,” “DU_SUAD,” and “DU_DD,” representing the duration until the hospital admission date from the ictus date, the duration until the stroke unit admission date from the hospital admission

Table 3
Variable description.

Feature	Description
ID_patient	Number of patient ID
Episode	Number of the Hospital Episode
Hospitalization_Date	Date of hospitalization in Stroke Unit
Entry_Date	Date of hospital entry
Entry_Place	Urgency, Intensive Care Unit, Observation Room
Days at Stroke Unit	Number of days at Stroke Unit
Exit Reason	Discharged, Other Units, Other Hospitals, Death
Exit Date	Exit Date
Hospitalization Time	Number of hospitalization days
ICTUS Date	Date of stroke episode
SDD_Gender	Patient's gender
SDD_Age	Patient's age
SDD_SII	Does the patient receive social integration income?
UM_DOAC	Does the patient take Direct Oral Anti-Coagulants?
UM_VKAs	Does the patient take Vitamin K Antagonists?
UM_Stopped taking medication	Does the patient stopped to take VKAs?
UM_PAI	Does the patient take Platelet Aggregation Inhibitor?
UM_Benzodiazepines	Does the patient take Benzodiazepines?
UM_PPI	Does the patient take Proton Pump Inhibitor?
UM_Statines	Does the patient take Statines?
UM_ Anti-hypertension	Does the patient take Anti-hypertension?
UM_Anti-Diabetic	Does the patient take Anti-Diabetic?
UM	Usual Medication
PSH_Smoking	Is the patient a smoker?
PSH_Hypertension	Does the patient have hypertension?
PSH_Stroke	Does the patient had already a previous stroke?
PSH_IHD	Does the patient had already an Ischemic Heart Disease?
PSH_T2DM	Does the patient have Type 2 diabetes mellitus?
PSH_AA	Is the patient alcoholic?
PSH_MH	Pre stroke medical history (other diseases or episodes)
PSH_Rankin	The modified Rankin scale level of the first stroke
PSH_AF	Does the patient have Atrial Fibrillation?
COA_BP	Blood pressure
COA_NIHSS	Patient's level according to the National Institutes of Health Stroke Scale
COA_territory	Basilar / Right / Left Middle Cerebral Artery
COA_ thrombolysis	Has the patient undergone thrombolysis?
COA_thrombectomy	Has the patient undergone thrombectomy?
SH_RI	Has the patient had Respiratory Infection during hospitalization?
SH_TC	Has the patient had Thrombotic Complications during hospitalization?
SH_UI	Has the patient had Urinary Infection during hospitalization?
SH_CIH	Are there any other complications during the hospitalization?
Rankin_2	Level of the Modified Rankin Scale after the stroke episode (target variable)
RHAD	Re-Hospitalization After Discharged?
BOA_RBC	Red Blood Cells
BOA_Hemoglobin	Hemoglobin
BOA_Hematocrit	Hematocrit
BOA_MCV	Mean Corpuscular Volume
BOA_MCH	Mean Corpuscular Hemoglobin
BOA_RDW	Red Cell Distribution Width
BOA_Leukocytes	Leukocytes
BOA_Neutrophils	Neutrophils
BOA_Lymphocytes	Lymphocytes
BOA_Monocytes	Monocytes
BOA_Eosinophils	Eosinophils
BOA_Basophils	Basophils
BOA_Platelets	Platelets
BOA_MPV	Mean Platelet Volume
BOA_PTC	Percutaneous transhepatic cholangiography
BOA_PDW	Platelet Distribution Width
BOA_PT	Prothrombin Time
BOA_INR	International Normalized Ratio
BOA_D_dimers	D-dimers
BOA_BNP	Brain natriuretic peptide test
BOA_Troponin_I	Troponin I
BOA_DCCT	Diabetes Control and Complications Trial
BOA_Glycaemia	Glycaemia
BOA_AST	Aspartate transaminase
BOA_ALT	Alanine transaminase
BOA_Na	Sodium
BOA_K	Potassium
BOA_ToC	Total Cholesterol
BOA_LDL	Low-Density Lipoprotein
BOA_HDL	High-Density Lipoprotein
BOA_triglycerides	Triglycerides
BOA_TP	Total Protein

(*continued on next page*)

Table 3 (*continued*)

Feature	Description
BOA_BUN	Blood Urine Nitrogen
BOA_Creatinine	Creatinine
BOA_UA	Uric Acid
BOA_CPA	Cardiopulmonary Arrest
BOA_FT4	Free Thyroxine Test
BOA_TSH	Thyroid Stimulating Hormone

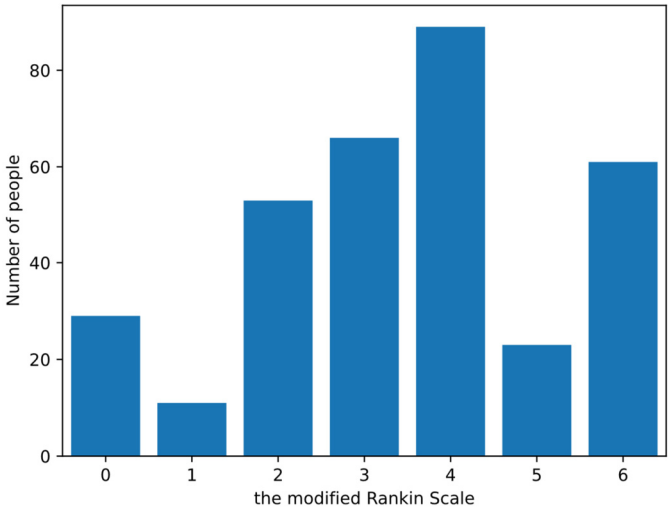


Fig. 2. The modified Rankin Scale class distribution.

date, and the duration until the discharge date from the stroke unit admission date, respectively.

Inconsistencies, writing errors, and values were corrected, and object columns were replaced with suitable int/float columns [55]. Binary variables were adjusted to a 1 or 0 format, and qualitative data values were grouped accordingly [71]. Two new columns, "COA_PB1" and "COA_PB2", were generated to store the first and second blood pressure values, respectively. Text columns were transformed, with NIHSS converted from categorical attributes to the proper numerical values, and the counts of usual medication pills and medical history were extracted and categorized [42]. Data quality was ensured by checking for duplicates and expected range values.

Finally, columns deemed unnecessary for analysis were dropped, including "Entry_Date", "Hospitalization_Date", "Exit_Date", "ICTUS_Date", "COA_BP", "UM", "PSH_MH", and "SH_CIH". Besides, "Exit_Reason" and "DU_DD" were discarded due to data leakage since one of the reasons for leaving is death, which is also the goal we want to predict the moment a person enters the hospital and not after days of hospitalization. Table 4 and Table 5 describe the 50 numerical and 27 categorical dependent features that remain after the feature engineering. Discrete variables are presented as counts (n) and percentages (%), and continuous variables are presented as mean and standard deviation [70].

To prepare the data for the modeling part, one hot encoding was performed for the categorical variables [71]. When dealing with missing values, 6 columns, and 10 rows were dropped because their percentage of missing values equaled or exceeded 40% [53]. For columns with a missing ratio lower than 0.4, the categorical values were filled in with the mode because there were few missing values, and the numerical values were filled in with K-Nearest Neighbors (KNN) imputer [53]. The choice of 30 nearest neighbors was taken because a small K may result in noise being introduced into the imputed values, while a large K may lead to over smoothing and loss of important patterns in the data [5].

Table 4
Statistics of numerical variables.

Feature	Mean	Standard Deviation
SDD_Age	72.6	12.4
PSH_Rankin	0.3	0.7
COA_NIHSS	14.8	6.4
BOA_RBC	4.6	0.5
BOA_Hemoglobin	136.0	17.1
BOA_Hematocrit	0.4	0.0
BOA_MCV	89.3	6.7
BOA_MCH	29.8	2.6
BOA_RDW	14.3	2.1
BOA_Leukocytes	9.1	3.1
BOA_Neutrophils	6.5	3.0
BOA_Lymphocytes	1.7	0.9
BOA_Monocytes	0.7	0.3
BOA_Eosinophils	0.1	0.1
BOA_Basophils	0.1	0.1
BOA_Platelets	211.7	64.3
BOA_MPV	9.7	1.4
BOA_PTC	0.2	0.1
BOA_PDW	16.3	1.8
BOA_PT	12.3	2.9
BOA_INR	1.1	0.3
BOA_D_dimers	2470.0	5693.9
BOA_BNP	320.5	561.3
BOA_Troponin_I	99.5	749.6
BOA_DCCT	6.1	1.4
BOA_Glycaemia	132.5	55.5
BOA_AST	33.3	23.9
BOA_ALT	24.8	23.0
BOA_Na	138.6	3.0
BOA_K	4.2	0.5
BOA_ToC	188.5	51.3
BOA_LDL	122.0	45.5
BOA_HDL	46.9	12.1
BOA_triglycerides	107.8	67.5
BOA_TP	6.7	0.7
BOA_BUN	21.8	10.1
BOA_Creatinine	1.0	0.5
BOA_UA	6.1	1.9
BOA_CPA	12.0	26.3
BOA_FT4	1.0	0.2
BOA_TSH	1.4	1.9
DU_HAD	0.2	0.8
DU_SUAD	0.0	0.1
COA_BP1	154.0	26.5
COA_BP2	84.0	16.1
UM_count	2.8	2.9
UM_category	1.4	1.0
PSH_MH_count	3.7	2.6
PSH_MH_category	2.2	0.9
SH_CIH_category	1.6	1.1

Data normalization was performed using Min-Max Scaler technique to ensure uniformity across features [69]. Two methods of dealing with outliers were tested [72]. One treatment was to bin all the numerical features into 3 bins according to the Decision Tree (DT) Regressor best-split criteria, making a model less sensitive to outliers. The choice of 3 bins is influenced by two important considerations. Firstly, an increasing number of bins potentially allows for a finer resolution in terms of scale and the possibility of identifying the relative change in the importance of risk factors over smaller intervals [20]. However, by increasing the number of bins, one reduces the statistical sample in each bin, thereby potentially losing statistical significance [20]. The other method was to change outliers' values, identified based on a threshold of 3 times the interquartile range to a random value from either 0 to 10% or 90% to 100% of the data accordingly [69]. Given the low univariance, exceptions were made for 'PSH_Rankin', 'DU_HAD', and 'DU_SUAD' to avoid null univariance of the variable.

Feature selection techniques help to find the optimal subset of relevant features, avoid overfitting, improve the prediction ac-

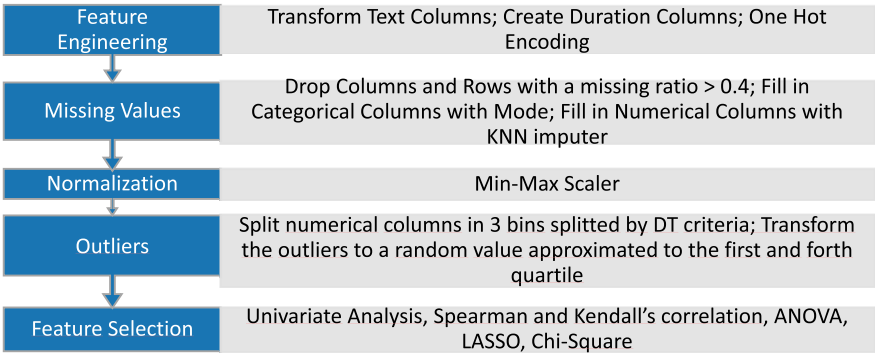


Fig. 3. Data preparation summary.

Table 5
Statistics of categorical variables.

Feature	Category	Count	Percentage
SDD_Gender	Male sex	193	58.1%
SDD_SII	Yes	91	27.7%
UM_DOAC	Yes	21	6.3%
UM_VKAs	Yes	18	5.4%
UM_Stopped_taking_medication	Yes	3	0.9%
UM_PAI	Yes	89	26.8%
UM_Benzodiaz	Yes	41	12.3%
UM_Statines	Yes	101	30.4%
UM_Anti_hypertension	Yes	192	57.8%
UM_Anti_Diabetic	Yes	53	16.0%
UM_PPI	Yes	58	17.5%
PSH_smoking	Yes	54	16.3%
PSH_hypertension	Yes	203	61.1%
PSH_stroke	Yes	50	15.1%
PSH_IHD	Yes	59	17.8%
PSH_T2DM	Yes	63	19.0%
PSH_AA	Yes	38	11.4%
PSH_AF	Yes	93	28.0%
COA_thrombolysis	Yes	175	53.0%
COA_thrombectomy	Yes	39	11.7%
SH_RI	Yes	104	31.5%
SH_TC	Yes	5	1.5%
SH_UI	Yes	59	17.8%
RHAD	Yes	65	21.6%
SH_CIH_binary	Yes	278	83.7%
Entry_Reason	Emergency	326	98.2%
Entry_Reason	Intensive Care/ Observation Room	6	1.8%
COA_territory	LMCA	172	51.8%
COA_territory	RMCA	144	43.4%
COA_territory	Basilar	16	4.8%

curacy of classifiers, and provide faster and more cost-effective models [69]. As so, multiple techniques were employed, including univariate analysis, analysis of variance (ANOVA), Spearman and Kendall's correlation, Chi-Square, and Least Absolute Shrinkage and Selection Operator (LASSO) [17,36,45,58]. Regarding ANOVA, Chi-Square, and Kendall's method, variables with a P-value of <0.05 were considered statistically significant [30].

The idea here is not to already identify risk factors for post-stroke mortality but rather to eliminate redundant and irrelevant variables that are clearly not candidates for risk factors [68]. Therefore, variables with little statistical evidence that have an absolute Spearman correlation value with other variables higher than 0.75 are discarded [25].

3.3. Modeling

Most scientific papers group in favorable and poor outcomes when predicting the mRS [21,30,65,70]. In this study, we are also predicting the mRS by grouping it as a binary classification problem where mRS $= 6$ is 1, and the other levels of the mRS are 0. But additionally, we are predicting the mRS both as a categorical target and as an ordinal regression to create a model capable of distinguishing the different levels.

The tenfold cross-validation was used to prevent overfitting, whereby 90% of cases are randomly selected as the training set and the remaining 10% as the testing set. This procedure is repeated five times to obtain the average results [70]. Alternatively to repeated stratified k-fold cross-validation, the leave one out method was used to provide a robust and unbiased estimation of the model performance given the small dataset by iteratively training the model on all but one data point and testing it on the omitted point [68].

For both classification problems, techniques such as LR, Gaussian Naïve Bayes, KNN, DT, Ridge classifier, Bagging, RF, Adaboost and Synthetic Minority Oversampling Technique (SMOTE) were used [14,16,20,28,29,56]. For the regression problem it was used Linear, Robust, Ridge, LASSO, Elastic Net, Stochastic Gradient Descent, RF, SVM Regressors and ANN [17].

Table 6 details the models and their inputs. All models were performed using version 1.1.2 of the Scikit Learn package, with the exception of ANN, which was created using version 2.12.0 of the TensorFlow package [1,51].

After testing different combinations of models, input variables, SMOTE, cross-validation, and outlier methods, the highest R-square was used to assess correlation in the regression problem [14,50]. For the classification problems, the highest Recall and F1 scores were obtained, leading to the automatic identification of risk variables of mortality post-stroke [38].

Besides ranking feature importance and checking binary features with high support, when analyzing risk factors, model-specific methods such as LionForests and model-agnostic methods such as SHAP were used to provide outcome explanations [18,36,43,44,54].

4. Results and discussion

This chapter presents the performance of the models developed in this study. Additionally, it identifies and analyzes risk factors both independently and collectively. Finally, it creates one general risk profile and discusses the variables and ML models in relation to those found in the literature review.

A total of 332 patients were registered during the study period. After excluding 10 (3%) patients with missing laboratory tests or clinical data, the mean age of the patients included was 72.8 $\pm$ 12.4 years and 57.8% were men. Of the 322 patients, 58 (18%) have died shortly after having a stroke.

The Random Forest (RF) model performed significantly better, when predicting mortality, than the remaining models, as it was

Table 6
Models and its inputs.

Model	Inputs
LogisticRegression()	penalty=['l2', 'l1', 'elasticnet', None], solver=['lbfgs', 'newton-cg', 'sag', 'saga'], max_iter=[100, 1000], tol=[0.1, 0.0001], C=[1, 10]
GaussianNB()	var_smoothing=[10, 1.5, 1, 0.5, 0.1, 0.01, 0.015, 0.001, 0.0001, 0.00001]
KNeighborsClassifier()	algorithm=['brute', 'auto'], n_neighbors=[1, 5, 10, 15, 16, 20, 25], p=[1,2], weights=['uniform', 'distance]
DecisionTreeClassifier()	max_depth=[5,6,7,8], criterion=['log_loss', 'entropy', 'gini']
RidgeClassifier()	'alpha'=[0.8], 'solver'=['svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga'], 'class_weight'=['balanced']
BaggingClassifier()	'base_estimator'=[RidgeClassifier(), LogisticRegression(), KNeighborsClassifier(n_neighbors=15), GaussianNB(var_smoothing=10)], 'bootstrap'=[True, False], 'bootstrap_features'=[True, False]
RandomForestClassifier()	'bootstrap'=[True, False], 'max_depth'=[2, 6, 10, 15, 20, 60, 80, None], 'max_features'=['sqrt', 'log2', None], 'n_estimators'=[40, 80, 100, 200, 300], 'min_samples_split'=[5, 10, 25, 50, 100, 200], warm_start=[True, False]
AdaBoostClassifier()	'base_estimator'=[RidgeClassifier(), Logistic Regression()], 'learning_rate'=[0.01, 0.1, 0.4, 0.7, 1], 'n_estimators'=[40, 150, 400]
LinearRegression()	fit_intercept=True, positive=False
RANSACRegressor()	min_samples=50, max_trials=100, loss='absolute_loss', residual_threshold=[10, 60]
HuberRegressor()	epsilon=7, alpha=0.9, warm_start=True
TheilSenRegressor()	fit_intercept=True, max_iter=300
Ridge()	solver='cholesky'
Lasso()	alpha=0.00001, positive=False
ElasticNet()	alpha=0.01, l1_ratio=0.4
SGDRegressor()	loss='squared_error', penalty='l2', alpha=0.0001, l1_ratio=0.15
ANN	4 middle dense layers with 32, 64, 128, 512 neurons respectively, ReLU activation function, Adam optimizer, loss='mse', epochs=100
RandomForestRegressor()	criterion='absolute_error', bootstrap=False, warm_start=[True,False], n_estimators=2000
SVR()	Kernel=['rbf', 'linear', 'poly', 'sigmoid', precomputed'], C=[1, 100], gamma=['auto', 0.1, 'scale'], coef0=[0, 1]

Table 7
Binary classification evaluation performance.

Model	F1	Precision	Accuracy	Recall
LogisticRegression()	55.8%	46.4%	79.2%	72.6%
GaussianNB()	48.1%	33.2%	64.7%	89.5%
KNeighborsClassifier()	54.2%	41.6%	75.3%	80.6%
DecisionTreeClassifier()	42.4%	34.1%	72.3%	59.1%
RidgeClassifier()	54.8%	43.9%	77.3%	76.1%
BaggingClassifier()	56.0%	47.9%	79.9%	70.8%
RandomForestClassifier()	94.6%	100.0%	98.5%	91.3%

Table 8
Multiclass classification evaluation performance.

Model	F1	Precision	Accuracy	Recall
LogisticRegression()	36.3%	36.3%	36.3%	36.3%
GaussianNB()	34.8%	34.8%	34.8%	34.8%
KNeighborsClassifier()	32.9%	32.9%	32.9%	32.9%
DecisionTreeClassifier()	30.1%	30.1%	30.1%	30.1%
RidgeClassifier()	27.1%	29.3%	29.7%	29.7%
BaggingClassifier()	33.2%	33.2%	33.2%	33.2%
RandomForestClassifier()	49.3%	50.2%	55.8%	55.8%
AdaBoostClassifier()	29.3%	29.4%	33.5%	33.5%

Table 9
Regression evaluation performance.

Model	MAE	MSE	RMSE	R-Square
LinearRegression()	1.082	1.992	1.407	33.8%
RANSACRegressor()	1.082	1.992	1.407	33.8%
HuberRegressor()	1.075	1.988	1.404	33.9%
TheilSenRegressor()	1.083	2.011	1.412	33.1%
Ridge()	1.082	1.990	1.406	33.9%
Lasso()	1.082	1.992	1.407	33.8%
ElasticNet()	1.073	1.987	1.404	34.0%
SGDRegressor()	1.074	1.989	1.405	33.9%
ANN	1.301	2.744	1.649	8.9%
RandomForestRegressor()	0.127	0.306	0.412	89.9%
SVR()	1.102	2.051	1.426	31.8%

Table 10
Death variables with support of 0.5 or more.

Feature	Support
SH_CIH_binary	0.948
ohc_BOA_ALT_bin_6.0-31.0	0.879
PSH_hypertension	0.690
ohc_BOA_PDW_bin_14.9-17.1	0.655
SH_RI	0.569
ohc_BOA_MCH_bin_28.8-31.4	0.500

found in [56] and [65], contrary to other studies where ANN had better results, probably due to the size of the dataset [30]. Nevertheless, RF was a disruptive model in predicting post-stroke mortality in Portugal, reaching 98.4% of accuracy and 91.4% of recall in the training set and 98.5% of accuracy and 91.3% of recall in the test set, as it shows Table 7 and Table 8. Tables 8 and 9 report metrics for multiclass classification and mRS regression for the test set, respectively.

The input variables that led to the best performance of the binary RF model are the risk factors of post-stroke mortality. They are 'SDD_Gender', 'SDD_SII', 'PSH_hypertension', 'COA_thrombolysis', 'COA_thrombectomy', 'SH_RI', 'RHAD', 'PSH_Rankin', 'COA_NIHSS', 'BOA_Hemoglobin', 'BOA_Hematocrit', 'BOA_MCH', 'BOA_RDW', 'BOA_Eosinophils', 'BOA_PDW', 'BOA_INR', 'BOA_Troponin_I', 'BOA_Glycaemia', 'BOA_ALT', 'BOA_Na', 'BOA_K', 'DU_HAD', 'UM_count', and 'SH_CIH_category'.

Fig. 4 shows the feature importance ranking. The most important feature when predicting the outcome was stroke severity level, as it was found in [2,6,9,11,16,19,22,24,29–31,45,47,56,57,63], and [70]. However, variables like age, AF, and smoking status were discarded from the risk factors, contrary to most scientific papers analyzed in the literature review [36].

The binary variables present in at least half of the deaths are listed in Table 10.

To understand the specific impact of individual features on post-stroke mortality, rank variables, identify outliers, analyze the spread and magnitude of SHAP values, and compare features, an overall SHAP feature plot is constructed and shown in Fig. 5. All factors are listed on the vertical axis ranked by SHAP importance [36]. X-axis represents the SHAP value, which indicates the degree of change in log odds. The color of each point on the graph represents the value of the corresponding numerical feature, with red indicating high values and blue indicating low values. For a specified binary factor, each point indicates a patient to whom that factor applies (in red) or does not apply (in blue) [36].

The right side of a patient in red means it has the impact to cause death, as it happens with the "COA_NIHSS", "BOA_Glycaemia", "SH_RI", "SH_CIH_category", and "DU_HAD" variables. This means that a higher value of stroke severity, glycemia, days from stroke event until hospital admission date, number of complica-

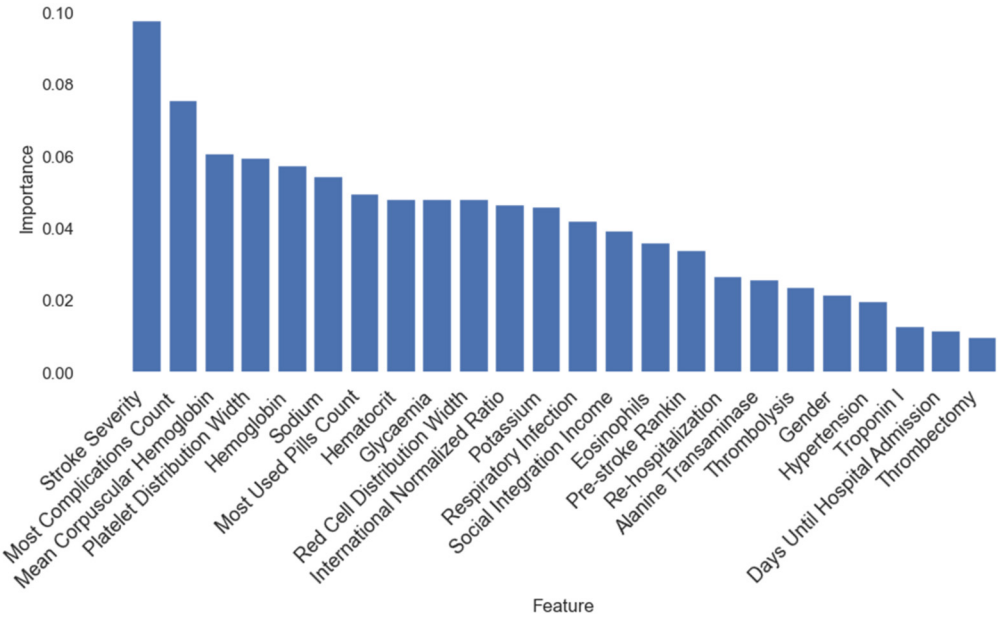


Fig. 4. Features importance.

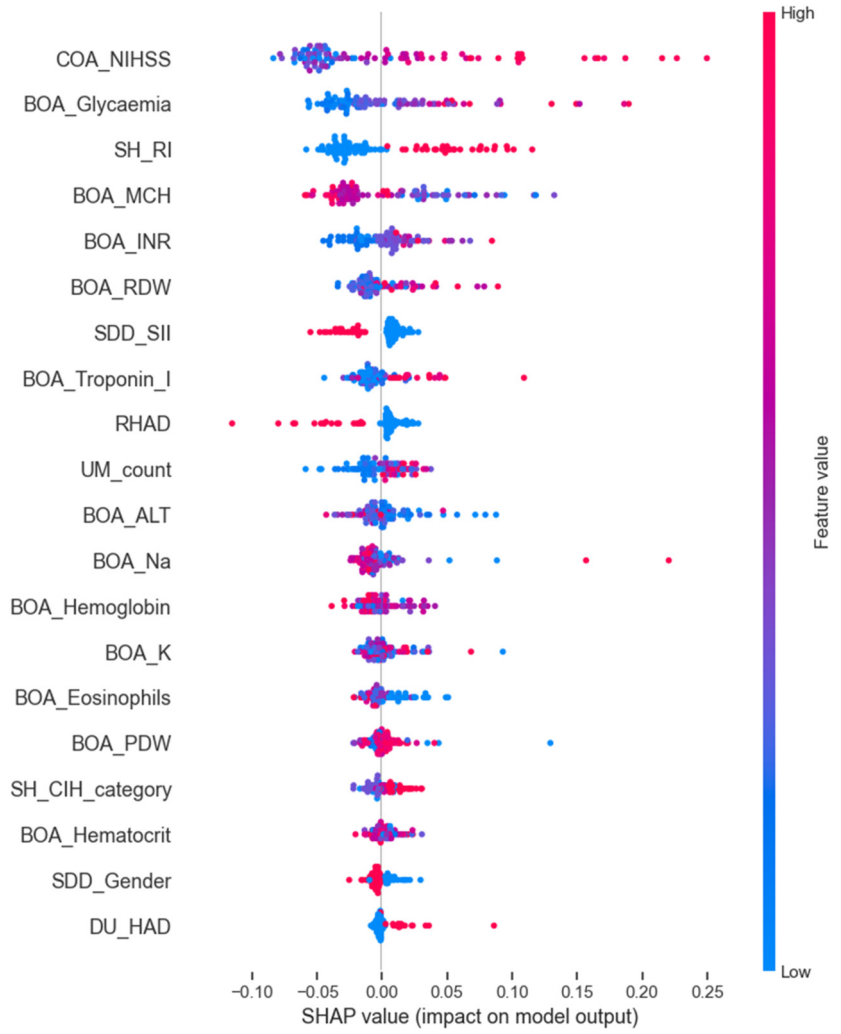


Fig. 5. SHAP values for feature importance.

Fig. 6. Importance ranking of first patient characteristics.

tions during the hospitalization, and the presence of respiratory infections are more likely to lead to mortality.

The opposite occurs with "RHAD" and "SDD_SII," where a higher value corresponds to a negative SHAP value. This means that people who were previously hospitalized or receive social income tend to negatively affect the output by staying alive. In addition, a SHAP value near 0 means that the corresponding factor makes a small contribution to the development of stroke [36].

After analyzing the entire dataset, we can examine it row by row to identify and characterize individual risk and healthy profiles. For example, Fig. 6 shows a force plot for a single patient who died after suffering a stroke. $E[f(x)] = 0.1791$ indicates the base value of shake of the overall sample [36]. The three most decisive factors contributing to death were, by order, the respiratory infection, the level of glycemia, and the alanine transaminase value at admission. On the other hand, despite not being strong enough, the red cell distribution width value had a negative impact, contributing to survival.

Finally, the SHAP value for the first patient is 0.74 (shown in bold in the upper right corner). This patient's illness is substantial compared with the value of E(x) [36]. Therefore, this individual meets the definition of post-stroke death.

In LionForests, the interpretations are presented in the form of rules [43]. Each rule is a conclusive set of conditions about the features that affect an instance's prediction [43]. LionForests implements feature and path reduction approaches to provide more minor rules containing conditions with broader ranges [43]. For example, the first patient's outcome can be explained by the following rule:

'if 15.5 <= COA_NIHSS <= 21.5 & 145.5 <= BOA_Glycaemia <= 538.0 & 31.65 <= BOA_MCH <= 38.1 & 4.002 <= BOA_K <= 4.052 & 13.8 <= BOA_RDW <= 14.95 & 135.5 <= BOA_Na <= 138.0 & 69.358 <= BOA_Troponin_I <= 421.142 & 147.5 <= BOA_Hemoglobin <= 156.0 & 6.0 <= BOA_ALT <= 17.383 & 0.465 <= BOA_Hematocrit <= 0.52 & 16.35 <= BOA_PDW <= 17.15 & 1.08 <= BOA_INR <=1.4 & 2.5 <= UM_count <= 5.5 & 0.5 <= SH_RI <= 1.0 & 0.0 <= RHAD <= 0.5 & 0.0 <= SDD_SII <= 0.5 then Dead'

This rule reduced the number of identified risk factors from 24 to 16. It also limited the interval range of the numerical features combined with the presence of a respiratory infection during hospitalization, the absence of a social integration income, and a re-hospitalization.

This study showcased the efficacy of ML models in accurately forecasting post-stroke mortality and elucidating its contributing factors, thereby achieving the set research goals. By delving deeper into explanatory techniques, it surpasses previous research [69].

5. Conclusion

Stroke ranks among the top causes of death, underscoring the pressing need to pinpoint and assess the factors contributing to mortality in stroke patients [41]. This study aimed to craft an effective ML model for predicting mortality and elucidating the impact of each factor, both individually and in combination with others.

RF was the model that outperformed, identifying the risk factors Alanine Transaminase, Counting Recurrent Complications, Counting Recurrent Pills, Days Until Hospital Admission, Eosinophils, Gender, Glycaemia, Hematocrit, Hemoglobin, Hypertension, International Normalized Ratio, Mean Corpuscular Hemoglobin, Platelet Distribution Width, Potassium, Pre-stroke Rankin, Red Cell Distribution Width, Re-hospitalization, Respiratory Infection, Sodium, Social Integration Income, Stroke Severity, Thrombectomy, Thrombolysis, and Troponin I values at admission.

Through this study, a comprehensive examination was conducted to understand the individual and collective impact of factors on stroke outcomes. This entailed establishing guidelines that delineate risk profiles and scrutinizing trends regarding the effect of altering variable values on outcomes. Three novel approaches were introduced to enhance post-stroke mortality analysis in Portugal. Firstly, it identifies prevalent variables in over fifty percent of fatalities, considering the two types of strokes: hemorrhagic and ischemic. Secondly, it explains straightforwardly how each risk factor influences mortality using SHAP. Lastly, it proposes rules that connect different factors to determine the outcome for each individual [36,43,44].

This study advances academia and research by enhancing healthcare predictive analytics with a precise mortality prediction ML model by identifying several risk factors, confirming stroke severity as the main one, and excluding other potential factors found in the literature review, such as age, AF, and smoking status, and by providing a more comprehensive analysis of risk factors than previous studies. This study also has implications for healthcare professionals, as it equips them with early identification and intervention tools for high-risk stroke patients. They can make informed decisions and create customized treatment plans, potentially improving clinical outcomes and reducing mortality.

This study is not without limitations. First, this is a single-center study of a single region with a small sample size, which limits generalization [46,64]. Second, despite internal cross-validation, the lack of external validation means overfitting cannot be ruled out [56]. Third, while ML can outperform traditional methods, their clinical implementation can be complex [56]. Fourth, the dataset has undergone various pre-processing steps like removing and imputing missing data, reducing the sample size even more, and introducing bias and uncertainty to interpretations. Last, the dataset is imbalanced, and RF tends to be biased toward the majority class because it aims to reduce overall impurity [7]. The splitting criteria will lead to suboptimal splits that do not adequately separate the minority class. As such, features distinguishing the majority class may seem overly important, while those distinguishing the minority class might be underrepresented. Future research should include multi-center and multi-region studies validated with independent datasets to enhance generalizability across different populations. Increasing sample sizes will provide more robust and reliable results, creating precise rules for defining high-risk profiles.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s) and/or volunteers.

Human and animal rights

The authors declare that the work described has not involved experimentation on humans or animals.

Funding

This work has been supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, X. Zheng, TensorFlow: a system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI'16), 2016, pp. 265–283.

[2] P. Abreu, R. Magalhães, D. Baptista, E. Azevedo, M.C. Silva, M. Correia, Readmissions and mortality during the first year after stroke—data from a population-based incidence study, Front. Neurol. 11 (July 2020) 1–11, https://doi.org/10.3389/fneur.2020.00636.

[3] A. Aigner, U. Grittner, A. Rolfs, B. Norrving, B. Siegerink, M.A. Busch, Contribution of established stroke risk factors to the burden of stroke in young adults, Stroke 48 (7) (2017) 1744–1751, https://doi.org/10.1161/STROKEAHA.117.016599.

[4] A. Alaiad, H. Najadat, B. Mohsen, K. Balhaf, Classification and association rule mining technique for predicting chronic kidney disease, J. Inf. Knowl. Manag. 19 (01) (2020) 2040015, https://doi.org/10.1142/S0219649220400158.

[5] T. Aljrees, Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning, PLoS ONE 19 (1 January 2024) 1–24, https://doi.org/10.1371/journal.pone.0295632.

[6] M.N. Andersen, K.K. Andersen, L.P. Kammersgaard, T.S. Olsen, Sex differences in stroke survival: 10-year follow-up of the Copenhagen stroke study cohort, J. Stroke Cerebrovasc. Dis. 14 (5) (2005) 215–220, https://doi.org/10.1016/j.jstrokecerebrovasdis.2005.06.002.

[7] M. Bader-El-Den, E. Teitei, T. Perry, Biased random forest for dealing with the class imbalance problem, IEEE Trans. Neural Netw. Learn. Syst. 30 (7) (2019) 2163–2172.

[8] J.P. Bembenek, K. Kurczych, B. Kłysz, A. Cudna, J. Antczak, A. Członkowska, Prediction of recovery and outcome using motor evoked potentials and brain derived neurotrophic factor in subacute stroke, J. Stroke Cerebrovasc. Dis. 29 (11) (2020) 1–7, https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105202.

[9] R. Bonita, M.A. Ford, A.W. Stewart, Predicting survival after stroke: a three-year follow-up, Stroke 19 (6) (1988) 669–673, https://doi.org/10.1161/01.STR.19.6.669.

[10] T. Bruening, M. Al-Khaled, Stroke-associated pneumonia in thrombolyzed patients: incidence and outcome, J. Stroke Cerebrovasc. Dis. 24 (8) (2015) 1724–1729, https://doi.org/10.1016/j.jstrokecerebrovasdis.2015.03.045.

[11] G. Çelik, Ö.K. Baykan, Y. Kara, H. Tireli, Predicting 10-day mortality in patients with strokes using neural networks and multivariate statistical methods, J. Stroke Cerebrovasc. Dis. 23 (6) (2014) 1506–1512, https://doi.org/10.1016/j.jstrokecerebrovasdis.2013.12.018.

[12] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0: step-by-step data mining guide, SPSS Inc 78 (2000) 1–78, http://www.crisp-dm.org/CRISPWP-0800.pdf.

[13] G. Chauhan, S. Debette, Genetic risk factors for ischemic and hemorrhagic stroke, Curr. Cardiol. Rep. 18 (12) (2016), https://doi.org/10.1007/s11886-016-0804-z.

[14] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (Sept. 28, 2002) 321–357, https://arxiv.org/pdf/1106.1813.pdf.

[15] L. Churilov, H. Johns, G. Turc, "Tournament methods" for the ordinal analysis of modified Rankin scale: the past, the present, and the future, Stroke 53 (10) (2022) 3032–3034, https://doi.org/10.1161/STROKEAHA.122.039614.

[16] M. Daidone, S. Ferrantelli, A. Tuttolomondo, Machine learning applications in stroke medicine: advancements, challenges, and future prospectives, Neural Regen. Res. 19 (4) (2024) 769–773, https://doi.org/10.4103/1673-5374.382228.

[17] S. Dev, H. Wang, C.S. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthc. Anal. 2 (February 2022) 100032, https://doi.org/10.1016/j.health.2022.100032.

[18] Y. Du, A.R. Rafferty, F.M. McAuliffe, C. Mooney, Explaining large-for-gestational-age births: a random forest classifier with a novel local interpretation method, in: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2021, 2021, p. 2020.

[19] T. Dziedzic, A. Slowik, J. Pera, A. Szczudlik, Beta-blockers reduce the risk of early death in ischemic stroke, J. Neurol. Sci. 252 (1) (2007) 53–56, https://doi.org/10.1016/j.jns.2006.10.007.

[20] J.F. Easton, C.R. Stephens, M. Angelova, Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach, Comput. Biol. Med. 54 (2014) 199–210, https://doi.org/10.1016/j.compbiomed.2014.09.003.

[21] A.K. Elhabr, J.M. Katz, J. Wang, M. Bastani, G. Martinez, M. Gribko, D.R. Hughes, P. Sanelli, Predicting 90-day modified Rankin scale score with discharge information in acute ischaemic stroke patients following treatment, BMJ Neurol. Open 3 (1) (2021) 1–9, https://doi.org/10.1136/bmjno-2021-000177.

[22] M. Eriksson, B. Norrving, A. Terént, B. Stegmayr, Functional outcome 3 months after stroke predicts long-term survival, Cerebrovasc. Dis. 25 (5) (2008) 423–429, https://doi.org/10.1159/000121343.

[23] M.C. Feng, Y.C. Lin, Y.H. Chang, C.H. Chen, H.C. Chiang, L.C. Huang, Y.H. Yang, C.H. Hung, The mortality and the risk of aspiration pneumonia related with dysphagia in stroke patients, J. Stroke Cerebrovasc. Dis. 28 (5) (2019) 1381–1387, https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.011.

[24] A. Fernandes, I. Taveira, R. Soares, A. Marreiros, H. Nzwalo, Impact of process of care in the short-term mortality in non-severe intracerebral hemorrhage in southern Portugal, J. Clin. Neurosci. 101 (February 2022) 259–263, https://doi.org/10.1016/j.jocn.2022.05.021.

[25] Y. Gao, Y. Wang, D. Li, J. Zhao, Z. Dong, J. zhou, G. Fu, J. Zhang, Disability assessment in stroke: relationship among the pictorial-based longshi scale, the barthel index, and the modified Rankin scale, Clin. Rehabil. 35 (4) (2021) 606–613, https://doi.org/10.1177/0269215520975922.

[26] GBD, Global burden of stroke disease in Portugal 2019, Global Burden of Disease Study 2019 (GBD 2019), https://vizhub.healthdata.org/gbd-compare/#, 2020.

[27] D.P. Ghelani, H.A. Kim, S.R. Zhang, G.R. Drummond, C.G. Sobey, T.M. De Silva, Ischemic stroke and infection: a brief update on mechanisms and potential therapies, Biochem. Pharmacol. 193 (July 2021) 114768, https://doi.org/10.1016/j.bcp.2021.114768.

[28] P. Govindarajan, R. KS, S. S, S. S, Impact of modifiable and non-modifiable risk factors on the prediction of stroke disease, in: International Conference on Trends in Electronics and Informatics ICEI 2017, 2017, pp. 985–989.

[29] H. Henon, O. Godefroy, D. Leys, F. Mounier-Vehier, C. Lucas, P. Rondepierre, A. Duhamel, J.P. Pruvo, Early predictors of death and disability after acute cerebral ischemic event, Stroke 26 (3) (1995) 392–398, https://doi.org/10.1161/01.STR.26.3.392.

[30] J.N. Heo, J.G. Yoon, H. Park, Y.D. Kim, H.S. Nam, J.H. Heo, Machine learning-based model for prediction of outcomes in acute stroke, Stroke 50 (5) (2019) 1263–1265, https://doi.org/10.1161/STROKEAHA.118.024293.

[31] K. Jamrozik, R.J. Broadhurst, S. Forbes, G.J. Hankey, C.S. Anderson, Predictors of death and vascular events in the elderly: the Perth community stroke study, Stroke 31 (4) (2000) 863–868, https://doi.org/10.1161/01.STR.31.4.863.

[32] R.R.A. Kadir, M. Alwjwaj, U. Bayraktutan, MicroRNA: an emerging predictive, diagnostic, prognostic and therapeutic strategy in ischaemic stroke, Cell. Mol. Neurobiol. 42 (5) (2022) 1301–1319, https://doi.org/10.1007/s10571-020-01028-5.

[33] F. Khan, S. Abusharha, A. Alfuraidy, K. Nimatallah, R. Almalki, R. Basaffar, M. Mirdad, M.F. Chevidikunnan, R. Basuodan, Prediction of factors affecting mobility in patients with stroke and finding the mediation effect of balance on mobility: a cross-sectional study, Int. J. Environ. Res. Public Health 19 (24) (2022), https://doi.org/10.3390/ijerph192416612.

[34] G. Khatri, M. Singh, S. Bika, K. Joshi, N. Swami, Carotid intima-media thickness: an independent risk factor for stroke prediction-a call for revised framingham score system, J. Anat. Soc. India 71 (3) (2022) 169–177, https://doi.org/10.4103/jasi.jasi_212_21.

[35] S. Lee, H. Lee, H.S. Kim, S.B. Koh, Incidence, risk factors, and prediction of myocardial infarction and stroke in farmers: a Korean nationwide population-based study, J. Prev. Med. Public Health 53 (5) (2020) 313–322, https://doi.org/10.3961/JPMPH.20.156.

[36] J. Li, Y. Luo, M. Dong, Y. Liang, X. Zhao, Y. Zhang, Z. Ge, Tree-based risk factor identification and stroke level prediction in stroke cohort study, BioMed Res. Int. 2023 (2023), https://doi.org/10.1155/2023/7352191.
[37] F. Liu, R.C. Tsang, J. Zhou, M. Zhou, F. Zha, J. Long, Y. Wang, Relationship of barthel index and its short form with the modified Rankin scale in acute stroke patients, J. Stroke Cerebrovasc. Dis. 29 (9) (2020) 105033, https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105033.
[38] F.R. Lucini, F.S. Fogliatto, G.J. Giovani, J.L. Neyeloff, M.J. de Anzanello, R.S. Kuchenbecker, B.D. Schaan, Text mining approach to predict hospital admissions using early medical records from the emergency department, Int. J. Med. Inform. 100 (2017) 1–8, https://doi.org/10.1016/j.ijmedinf.2017.01.001.
[39] I.L. Maier, A. Karch, R. Mikolajczyk, M. Bähr, J. Liman, Effect of beta-blocker therapy on the risk of infections and death after acute stroke - a historical cohort study, PLoS ONE 10 (2) (2015) 1–10, https://doi.org/10.1371/journal.pone.0116836.
[40] B. Manwani, C. Finger, L. Lisabeth, Strategies for maintaining brain health: the role of stroke risk factors unique to elderly women, Stroke 53 (8) (2022) 2662–2672, https://doi.org/10.1161/STROKEAHA.121.036894.
[41] S. Mateus, Capítulo I Acidente vascular cerebral: Definição epidemiologia e caraterização, Acidente Vasc. Cereb. (2015) 1–38, https://dspace.uevora.pt/rdpc/bitstream/10174/16842/13/11%C2%BA%20Cap%C3%ADtulo%20I%20AVC%20Defini%C3%A7%C3%A3o,%20epidemiologia%20e%20carateriza%C3%A7%C3%A3o%20final%208.pdf.
[42] D.K. Mishra, J. Kumar, J.K. Chaudhary, D.A.K. Upadhyay, D.S. Sharma, Role of text mining to enhance the quality of product using an unsupervised machine learning approach, ECS Trans. 107 (1) (2022) 12553–12560, https://doi.org/10.1149/10701.12553ecst.
[43] I. Mollas, N. Bassiliades, G. Tsoumakas, Conclusive local interpretation rules for random forests, Data Min. Knowl. Discov. 36 (4) (2022) 1521–1574, https://doi.org/10.1007/s10618-022-00839-y.
[44] I. Mollas, N. Bassiliades, I. Vlahavas, G. Tsoumakas, LionForests: local interpretation of random forests, CEUR Workshop Proc. 2659 (2020) 17–24.
[45] K. Nedeltchev, N. Renz, A. Karameshev, T. Haefeli, C. Brekenfeld, N. Meier, L. Remonda, G. Schroth, M. Arnold, H.P. Mattle, Predictors of early mortality after acute ischaemic stroke, Swiss Med. Wkly. 140 (17–18) (2010) 254–259.
[46] H. Nzwalo, C. Félix, J. Nogueira, P. Guilherme, F. Ferreira, T. Salero, S. Ramalhete, J. Martinez, M. Mouzinho, A. Marreiros, L. Thomassen, N. Logallo, Predictors of long-term survival after spontaneous intracerebral hemorrhage in southern Portugal: a retrospective study of a community representative population, J. Neurol. Sci. 394 (July 2018) 122–126, https://doi.org/10.1016/j.jns.2018.09.019.
[47] H. Nzwalo, J. Nogueira, A.C. Félix, P. Guilherme, P. Abreu, T. Figueiredo, F. Ferreira, A. Marreiros, L. Thomassen, N. Logallo, Short-term outcome of spontaneous intracerebral hemorrhage in Algarve, Portugal: retrospective hospital-based study, J. Stroke Cerebrovasc. Dis. 27 (6) (2018) 1721, https://doi.org/10.1016/j.jstrokecerebrovasdis.2018.01.016.
[48] M.J. O'Donnell, S.L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang, P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay, P. Lopez-Jaramillo, A. Damasceno, P. Langhorne, M.J. McQueen, A. Rosengren, M. Dehghan, G.J. Hankey, A.L. Dans, A. Elsayed, S. Yusuf, Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study, Lancet 388 (10046) (2016) 761–775, https://doi.org/10.1016/S0140-6736(16)30506-2.
[49] M.J. O'Donnell, X. Denis, L. Liu, H. Zhang, S.L. Chin, P. Rao-Melacini, S. Rangarajan, S. Islam, P. Pais, M.J. McQueen, C. Mondo, A. Damasceno, P. Lopez-Jaramillo, G.J. Hankey, A.L. Dans, K. Yusoff, T. Truelsen, H.C. Diener, R.L. Sacco, S. Yusuf, Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study, Lancet 376 (9735) (2010) 112–123, https://doi.org/10.1016/S0140-6736(10)60834-3.
[50] V. Papavasileiou, H. Milionis, C.J. Smith, K. Makaritsis, B.D. Bray, P. Michel, E. Manios, K. Vemmos, G. Ntaios, External validation of the prestroke independence, sex, age, national institutes of health stroke scale (ISAN) score for predicting stroke-associated pneumonia in the Athens stroke registry, J. Stroke Cerebrovasc. Dis. 24 (11) (2015) 2619–2624, https://doi.org/10.1016/j.jstrokecerebrovasdis.2015.07.017.
[51] 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, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[52] S. Pereira, F.B. Coelho, H. Barros, Acidente vascular cerebral: Hospitalização, mortalidade e prognóstico, Acta Med. Port. 17 (3) (2004) 187–192.
[53] L. Ren, T. Wang, A. Sekhari Seklouli, H. Zhang, A. Bouras, A review on missing values for main challenges and methods, Inf. Sci. 119 (2023) 102268, https://doi.org/10.1016/j.is.2023.102268.
[54] D. Salami, C.A. Sousa, Martins M. do R. O., C. Capinha, Predicting Dengue importation into Europe, using machine learning and model-agnostic methods, Sci. Rep. 10 (1) (2020) 1–13, https://doi.org/10.1038/s41598-020-66650-1.
[55] J.S. Saltz, CRISP-DM for data science: strengths, weaknesses and potential next steps, in: Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021, 2021, pp. 2337–2344.
[56] D. Scrutinio, C. Ricciardi, L. Donisi, E. Losavio, P. Battista, P. Guida, M. Cesarelli, G. Pagano, G. D'Addio, Machine learning to predict mortality after rehabilitation among patients with severe stroke, Sci. Rep. 10 (1) (2020) 1–10, https://doi.org/10.1038/s41598-020-77243-3.
[57] R.M. Shavelle, J.C. Brooks, D.J. Strauss, L. Turner-Stokes, Life expectancy after stroke based on age, sex, and Rankin grade of disability: a synthesis, J. Stroke Cerebrovasc. Dis. 28 (12) (2019) 104450, https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.104450.
[58] R. Shi, T. Zhang, H. Sun, F. Hu, Establishment of clinical prediction model based on the study of risk factors of stroke in patients with type 2 diabetes mellitus, Front. Neuroendocrinol. 11 (August 2020) 1–14, https://doi.org/10.3389/fendo.2020.00559.
[59] Y. Shinya, S. Miyawaki, I. Kumagai, T. Sugiyama, A. Takenobu, N. Saito, A. Teraoka, Risk factors and outcomes of cerebral stroke in end-stage renal disease patients receiving hemodialysis, J. Stroke Cerebrovasc. Dis. 29 (4) (2020) 104657, https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.104657.
[60] S. Das, A. Adhikary, A.A. Laghari, S. Mitra, Eldo-care: EEG with Kinect sensor based telehealthcare for the disabled and the elderly, Neurosci. Inform. 3 (2) (2023) 100130, https://doi.org/10.1016/j.neuri.2023.100130.
[61] K.B. Slot, E. Berge, P. Dorman, S. Lewis, M. Dennis, P. Sandercock, Impact of functional status at six months on long term survival in patients with ischaemic stroke: prospective cohort studies, BMJ 336 (7640) (2008) 376–379, https://doi.org/10.1136/bmj.39456.688333.BE.
[62] R. Soares, A. Fernandes, I. Taveira, A. Marreiros, H. Nzwalo, Predictors of pneumonia in patients with acute spontaneous intracerebral hemorrhage in Algarve, Southern Portugal, Clin. Neurol. Neursurg. 221 (July 2022), https://doi.org/10.1016/j.clineuro.2022.107387.
[63] J. Teles, J. Martinez, M. Mouzinho, P. Guilherme, A. Marreiros, H. Nzwalo, Gender differences in long-term mortality after spontaneous intracerebral hemorrhage in southern Portugal, Porto Biomed. J. 6 (4) (2021) e137, https://doi.org/10.1097/j.pbj.0000000000000137.
[64] B. Wang, Y. Xu, P. Wan, S. Shao, F. Zhang, X. Shao, J. Wang, Y. Wang, Right atrial fluorodeoxyglucose uptake is a risk factor for stroke and improves prediction of stroke above the CHA2DS2-VASc score in patients with atrial fibrillation, Front. Cardiovasc. Med. 9 (July 2022) 1–13, https://doi.org/10.3389/fcvm.2022.862000.
[65] H.L. Wang, W.Y. Hsu, M.H. Lee, H.H. Weng, S.W. Chang, J.T. Yang, Y.H. Tsai, Automatic machine-learning-based outcome prediction in patients with primary intracerebral hemorrhage, Front. Neurol. 10 (AUG 2019) 1–7, https://doi.org/10.3389/fneur.2019.00910.
[66] M. Wang, S.S. Rajan, A.P. Jacob, N. Singh, S.A. Parker, R. Bowry, J.C. Grotta, J.M. Yamal, Retrospective collection of 90-day modified Rankin scale is accurate, Clin. Trials 17 (6) (2020) 637–643, https://doi.org/10.1177/1740774520942466.
[67] Q. Wang, Y. Liu, L. Han, F. He, N. Cai, Q. Zhang, J. Wang, Risk factors for acute stroke-associated pneumonia and prediction of neutrophil-to-lymphocyte ratios, Am. J. Emerg. Med. 41 (2021) 55–59, https://doi.org/10.1016/j.ajem.2020.12.036.
[68] T.T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognit. 48 (9) (2015) 2839–2846, https://doi.org/10.1016/j.patcog.2015.03.009.
[69] J. Wu, Q. Zhang, L. Tao, X. Lu, Influencing factors analysis and prediction model development of stroke: the machine learning approach, J. Inf. Knowl. Manag. 22 (1) (2023) 1–16, https://doi.org/10.1142/S0219649222500794.
[70] T.C. Wu, Y.L. Liu, J.H. Chen, C.H. Ho, Y. Zhang, M.Y. Su, Prediction of poor outcome in stroke patients using radiomics analysis of intraparenchymal and intraventricular hemorrhage and clinical factors, Neurol. Sci. 44 (4) (2023) 1289–1300, https://doi.org/10.1007/s10072-022-06528-4.
[71] C.C. Yang, O.A. Bamodu, L. Chan, J.H. Chen, C.T. Hong, Y.T. Huang, C.C. Chung, Risk factor identification and prediction models for prolonged length of stay in hospital after acute ischemic stroke using artificial neural networks, Front. Neurol. 14 (2023), https://doi.org/10.3389/fneur.2023.1085178.
[72] J. Yang, L. Ji, Q. Wang, X. Lu, The prediction model of stroke on climate factors by multiple regression, in: Proceedings of 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2016, 2016, pp. 587–591.
[73] K. Yi, Y. Inatomi, M. Nakajima, T. Yonehara, M. Ueda, Reliability of the modified Rankin scale assessment using a simplified questionnaire in Japanese, J. Stroke Cerebrovasc. Dis. 30 (2) (2021) 105517, https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105517.
[74] Z. Yuanyuan, W. Jiaman, Q. Yimin, Y. Haibo, Y. Weiqu, Y. Zhuoxin, Comparison of prediction models based on risk factors and retinal characteristics associated with recurrence one year after ischemic stroke, J. Stroke Cerebrovasc. Dis. 29 (4) (2020) 104581, https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.104581.
[75] C. Zhang, X. Zhao, C. Wang, L. Liu, Y. Ding, F. Akbary, Y. Pu, X. Zou, W. Du, J. Jing, Y. Pan, K.S. Wong, Y. Wang, Y. Wang, Prediction factors of recurrent ischemic events in one year after minor stroke, PLoS ONE 10 (3) (2015) 1–12, https://doi.org/10.1371/journal.pone.0120105.
[76] X. Zhang, L. Xiao, L. Niu, Y. Tian, K. Chen, Comparison of six risk scores for stroke-associated pneumonia in patients with acute ischemic stroke: a systematic review and Bayesian network meta-analysis, Front. Med. 9 (2022), https://doi.org/10.3389/fmed.2022.964616.
[77] M. Zhou, X. Liu, F. Zha, F. Liu, J. Zhou, M. Huang, W. Luo, W. Li, Y. Chen, S. Qu, K. Xue, W. Fu, Y. Wang, Stroke outcome assessment: optimizing cutoff scores

for the longshi scale, modified Rankin scale and barthel index, PLoS ONE 16 (5 May 2021) 1–13, https://doi.org/10.1371/journal.pone.0251103.

[78] Y. Zhuo, H. Yu, Z. Yang, B. Zee, J. Lee, L. Kuang, Prediction factors of recurrent stroke among Chinese adults using retinal vasculature characteristics, J. Stroke Cerebrovasc. Dis. 26 (4) (2017) 679–685, https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.01.020.